



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

공학석사학위논문

마코프랜덤필드를 이용한
설문지에서의 비정상 반응 패턴 탐지

2015년 2월

서울대학교 대학원
협동과정 인지과학 전공
김 형 준

마코프랜덤필드를 이용한 설문지에서의 비정상 반응 패턴 탐지

지도교수 김 청 택

지도교수 장 병 탁

이 논문을 공학석사 학위논문으로 제출함

2015 년 2 월

서울대학교 대학원

협동과정:인지과학

김 형 준

김형준의 공학석사 학위논문을 인준함

2015 년 2 월

위 원 장 _____ 박 주 용 (인)

부위원장 _____ 장 병 탁 (인)

위 원 _____ 김 청 택 (인)

국문초록

본 연구는 마코프랜덤필드(MRFs)를 이용하여 설문지에서 무작위 반응과 같은 비정상 반응 패턴을 탐지하는데 주 목적이 있다. MRFs의 포텐셜 함수는 1) 특정 문항의 속성과 2) 이전 문항의 반응과 현재 문항의 반응에 대한 전이를 결합하여 구성하였다.

연구 1에서는 MRFs 모형이 비정상 반응 패턴을 잘 탐지할 수 있는지를 알아보기 위해 ROC curve의 AUC(Area Under the Curve)로 MMPI-2의 왜곡지표(VRIN, TRIN)와 IRT, SVM의 모형을 비교하였다. 그 이후 비정상 반응 패턴을 잘 판별할 수 있는 MRFs의 기준을 두 가지로 선정하였다. 1) 표준화된 개인의 로그 우도에서의 -2 와 2) 모든 문항에 대하여 무선적으로 반응한 반응 패턴의 로그 우도를 기준으로 하여 왜곡지표(VRIN, TRIN) 및 IRT 모형과의 정확도를 비교하였다. 연구 1의 결과 AUC와 정확도 모두에서 MRFs 모형이 다른 모형에 비해 여러 조건들에서 일관되게 비정상 반응 패턴을 잘 탐지하였다.

연구 2에서는 MRFs 모형을 활용하여 설문지 길이에 따른 비정상 반응 패턴의 증가를 알아보았다. 이를 위해 첫째, 설문지를 이등분, 그리고 삼등분하여 각 구간에서 비정상 반응 패턴이 증가하는 정도를 알아보았다. 둘째, 이동창(moving window)을 이용하여 설문 길이 증가에 따른 비정상 반응 패턴의 증가를 연구하였다. 연구 2의 결과 20분이 소요되는 짧은 설문지에서는 시간이 지남에 따라 비정상 반응 패턴이 증가하지 않았지만, 50분이 소요되는 비교적 긴 설문지에서는 시간이 지남에 따라 비정상 반응 패턴이 증가함을 보였다.

주요어 : 마코프랜덤필드, 비정상 반응 탐지, 반응 패턴 탐지

학 번 : 2012-22914

목 차

1. 서론	1
1.1 연구 배경	1
1.2 논문의 구성	5
2. 선행연구	6
2.1. MMPI-2 타당도 척도	6
2.2. 문항반응이론에 근거한 개인적합도	7
2.2.1. 문항반응이론	7
2.2.2. 문항반응이론의 가정 및 문제점	9
2.2.3. 문항반응이론을 이용한 개인적합도	11
3. 마코프랜덤필드 모형	13
3.1. 마코프랜덤필드	13
3.2. 마코프랜덤필드의 가정 및 결합확률	15
3.3. 이진 마코프랜덤필드 모형	16
3.3.1. 모형 상세화	17
3.3.2. 파라미터 추정법 및 개인적합도	20
4. 모형 평가 지표	25
5. 연구 1	29
5.1. 설문자료 및 실험방법	29
5.2. 실험결과	31

5.2.1. AUC를 이용한 모형 평가	31
5.2.2. 마코프랜덤필드 모형에서의 컷오프 기준	40
5.2.3. 컷오프 기준을 이용한 모형 평가	45
 6. 연구 2	 52
6.1. 설문자료 및 자료분석방법	53
6.1.1. 설문자료	53
6.1.2. 자료분석방법	54
6.2. 분석결과 및 논의	55
 7. 결론	 60
 참고문헌	 62

표 목 차

표 1: 설문지 구성 예시	18
표 2: ψ_i (노드 포텐셜)	21
표 3: ‘정이행 관계’의 $\psi_{i,i+1}$ (엣지 포텐셜)	22
표 4: ‘역이행 관계’의 $\psi_{i,i+1}$ (엣지 포텐셜)	22
표 5: 모든 반응 패턴들에 대한 포텐셜 값	23
표 6: 모든 반응 패턴들에 대한 확률	24
표 7: Confusion Matrix 예시	25
표 8: VRIN, TRIN, IRT, MRFs, SVM의 평균 AUC(총 567문항)	33
표 9: IRT, MRFs, SVM의 평균 AUC(총 30문항)	34
표 10: 모의 실험 1. 무선 문항 및 무선 반응 조작	36
표 11: 모의 실험 2. 연속 문항 및 연속 반응 조작	37
표 12: VRIN의 Confusion Matrix	45
표 13: TRIN의 Confusion Matrix	45
표 14: IRT(l_z)의 Confusion Matrix	45
표 15: IRT(l_0)의 Confusion Matrix	45
표 16: MRFs의 Confusion Matrix	46
표 17: VRIN, TRIN, IRT, MRFs의 평균 ACC(총 567문항)	47
표 18: VRIN, TRIN, IRT, MRFs의 평균 B. ACC(총 567문항)	48
표 19: 비정상 반응 패턴 비율에 따른 VRIN, TRIN, IRT, MRFs의 평균 ACC (총 567문항)	49
표 20: 비정상 반응 패턴 비율에 따른 VRIN, TRIN, IRT, MRFs의 평균 B. ACC (총 567문항)	49

표 21: 비정상 반응 패턴 비율에 따른 IRT와 MRFs의 평균 ACC 및 B. ACC (총 30문항)	50
표 22: 짧은 설문지에서의 전반부, 후반부의 비정상 반응 패턴의 수	55
표 23: 짧은 설문지에서의 초반부, 중반부, 종반부의 비정상 반응 패턴의 수	55
표 24: 긴 설문지에서의 전반부, 후반부의 비정상 반응 패턴의 수	57
표 25: 긴 설문지에서의 초반부, 중반부, 종반부의 비정상 반응 패턴의 수	57

그 립 목 차

그림 1: 2모수 로지스틱 모형(2PL)의 ICC	10
그림 2: 마코프 체인 모형	14
그림 3: 1차 이웃 시스템	14
그림 4: 2차 이웃 시스템	14
그림 5: N차 이웃 시스템	14
그림 6: 모든 사람들의 반응 패턴	17
그림 7: 연구 모형(반응 패턴 모형)	17
그림 8: MRFs 개인적합도 히스토그램	24
그림 9: ROC curve 예제	27
그림 10: VRIN, TRIN, MRFs, IRT의 ROC Curve	31
그림 11: 모의 실험 1(무선 문항 및 무선 반응 조작) 결과	38
그림 12: 모의 실험 2(연속 문항 및 연속 반응 조작) 결과	38
그림 13: 컷오프 기준 변화에 따른 ACC (총 567문항)	41
그림 14: 컷오프 기준 변화에 따른 ACC (총 30문항)	42
그림 15: 모의실험1의 ACC	43
그림 16: 모의실험2의 ACC	44
그림 17: 정상 반응 패턴과 비정상 반응 패턴들에 대한 MRFs 모형에서의 로그 우도 히스토그램	46
그림 18: 비정상 반응 패턴 비율에 따른 IRT, MRFs, SVM, TRIN, VRIN의 ACC 및 B. ACC. (총 567문항)	51
그림 19: 짧은 설문지에서 문항 위치에 따른 비정상 반응 패턴 수	56
그림 20: 긴 설문지에서 문항 위치에 따른 비정상 반응 패턴 수	58

1. 서론

1.1 연구 배경

설문조사는 설문지(questionnaire)에 대한 반응 패턴(반응의 시퀀스)을 통해 개인의 고유한 특성(태도, 성격, 지능 등)을 파악하는 것이다. 그러나, 연구자의 기대와 달리 사람들은 거짓 보고, 피로감 혹은 무관심 등으로 불성실한 응답을 보인다. 설문조사에서 적게는 3%에서 많게는 50% 정도의 사람들이 전체 문항 혹은 일부 문항에 대하여 불성실한 응답을 보인다(Curran et al., 2010; Ehlers et al., 2009).

즉, 설문조사와 같은 인간 데이터의 경우 많은 잡음(noise)이 존재할 수 있다. 그러나, 데이터가 많거나 설문 구성이 복잡한 설문지의 경우 사람이 일일이 모든 사람들의 반응 패턴을 보면서 불성실한 응답으로 분류하기가 쉽지 않다.

통계학적인 기법을 통하여 불성실한 반응으로 인한 오차를 추정하려는 시도는 있었지만, 그 특성에 따라 추정하기 비교적 용이한 오차와 그렇지 않은 오차가 발견되었다. Weisberg(2009)는 설문조사에서의 오차를 크게 표집오차와 비표집오차로 구분하고 있다. 통계학적으로 표집오차는 추정할 수 있고 샘플 사이즈를 증가시킴으로써 통제할 수 있지만, 비표집오차는 추정하기도 어렵고 통제하기도 어렵다. 이 중 불성실한 반응 패턴은 비표집오차에 속하므로 별도의 문항 혹은 모형을 통해 이를 탐지하여 제거하는 것이 필요하다.

Nichols et al. (1989)는 설문조사에서 불성실한 반응 패턴을 두 종류로 구분하고 있다. 첫 번째는 자신의 고유한 특성을 속임(faking)으로서 자신을 더 좋은 사람으로 보이게 하려 반응하는 것이고, 두 번째는 설문지의 특성과 관계없이 무작위로(randomly) 반응하는 것이다.

본 논문에서는 두 번째 특성(무작위 반응)을 지닌 사람들의 반응 패턴을 비정상(aberrant) 반응 패턴으로 명명하고 이를 탐지하는 모형을 제안하고자 한다. 장은경(2010)은 비정상 반응 패턴을 좀 더 정교화하여 1) 무선반응(random response) 2) 묵종반응(acquiescence response) 3) 비묵종반응(nonacquiescence response)으로 구분하였다. 1) 무선반응은 문항의 내용을 고려하지 않고 비일관적으로 응답하는 반응경향을 말하며, 2) 묵종반응과 3) 비묵종반응의 경우 여러 답안(alternative) 중에서 특정 선택지(“그렇다(1)” 혹은 “아니다(0)”)만을 한 줄로 응답하는 것을 말한다(Dahlstrom, et al., 1972).

Butcher et al. (1985, 1989)는 설문지에 무작위 반응 패턴을 탐지할 수 있는 척도를 포함시켰다. 그들은 설문지 Minnesota Multiphasic Personality Inventory-2(MMPI-2; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1985, 1989)를 제작하는 과정에서 두 개의 비일관성 척도(batteries)를 별도 문항들로 구성하였다. 첫 번째는 무선반응 비일관성 척도(Variable Response Inconsistency, VRIN)이고, 두 번째는 고정반응 비일관성 척도(True Response Inconsistency, TRIN)이다. 장은경(2010)의 분류기준으로 보면, VRIN은 무선반응을 탐지하기 위해 제작되었고 TRIN은 묵종반응과 비묵종반응을 탐지하기 위해 제작되었다. 그러나, MMPI-2의 두 왜곡지표(VRIN, TRIN)는 총 111 문항으로서 연구 참여자의 부담이 매우 크다. 따라서, 별도의 문항 없이도 통계 모형을 통해 비정상 반응 패턴들을 분류할 필요가 있다.

설문지에서 나타나는 비정상 반응 패턴들을 판별하기 위한 노력은 심리학 및 교육학에서 있어왔다. 여러 종류의 불성실한 반응 패턴을 탐지하기 위해 모수적(parametric) 추정 모형과 비모수적(non-parametric) 추정 모형을 합쳐 36종류의 통계치들이 존재한다(Karabatos, 2003). 그 중 문항반응이론(Item Response Theory, IRT)에 근거한 개인별 우도(l_0)를 사용한 연구가 가장 활발하고 여러 도메인(능력, 성격 등의 특성)에서 가

장 일관된 결과를 보인다(Reise & Due, 1991).

그러나, 문항반응이론의 경우 비정상 반응 패턴을 판별하기 위한 모형이 아니라 사람들의 특성과 문항의 특성을 추정하기 위한 모형이다. 즉, 이 모형은 비정상 반응 패턴이 없다는 것을 가정하고 있다. 따라서, 비정상 반응 패턴들이 많을 경우 모형의 파라미터(parameter)들을 불안정하게 추정할 수 있다.

한편, 기계학습(machine learning) 관점에서 지지벡터머신(support vector machine, SVM) 등의 감독학습(supervised learning)은 분류 성능은 우수할 수 있으나, 실제 비정상 반응 패턴들을 탐지하는 것에 활용하기는 어렵다. 왜냐하면, 실제 데이터에서는 ‘정상 반응’ 혹은 ‘비정상 반응’ 등의 라벨이 존재하지 않기 때문이다. 또한, 무감독학습(unsupervised learning)에서 k-means 등의 클러스터링(clustering)은 실제로 데이터 중에 비정상 반응 패턴이 존재하지 않더라도 비정상 반응 패턴으로 오분류(miss classification) 할 수 있다. 또한, 분류하더라도 어느 집단(group)이 정상 반응 집단이고 비정상 반응 집단인지 파악하기 힘들다는 단점이 있다.

따라서, 특정 반응 패턴에 대하여 무감독학습을 하면서도 특정 값을 기준으로 하여 정상 반응 패턴과 비정상 반응 패턴을 판별해야 한다. 마코프랜덤필드(markov random fields, MRFs)는 확률 그래프 모형으로서 무감독학습을 하면서도 주어진 모형에서 사람들의 반응 패턴에 대한 확률을 추정할 수 있다. 잡음이 섞인 이미지에서 잡음이 없는(noise-free) 이미지를 디코딩(decoding) 하듯이, 잡음이 섞인 사람들의 반응 패턴들 역시 패턴들 중 잡음이 없는 가장 전형적인 반응 패턴을 디코딩 할 수 있다. 단, 본 연구의 목적상 가장 확률이 높은 반응 패턴을 찾는 것이 아니라, 역으로 확률이 낮은 반응 패턴들을 찾는 것이다.

MRFs는 오류 함수(error function)를 연구자가 비교적 쉽게 정의할 수 있다. 즉, MMPI-2에서 두 종류의 왜곡 지표(VRIN, TRIN) 특성을 동시에 모형에 반영할 수 있다. MRFs는 무방향 그래프로서 $G=(X,E)$

로 표현 할 수 있다. 즉, MRFs는 노드(X , node)와 노드 간 연결을 나타내는 엣지(E , edge)로 모형을 나타낼 수 있다. 문항(X_i)을 노드(node 혹은 vertex)라고 하였을 때, 각 문항의 속성은 노드 포텐셜(node potential)에 반영할 수 있고, 이전 문항과 현재 문항의 전이는 엣지 포텐셜(edge potential)에 반영할 수 있다.

우선, 연구 1에서는 MMPI-2 설문지를 사용 하였다. 왜냐하면, 별도의 왜곡지표(문항들) 없이도 비정상 반응 패턴을 더 잘 탐지할 수 있음을 알아보기 위해 기존 왜곡지표(VRIN, TIRN)와 MRFs 모형의 판별 정확도를 비교해야 한다. MRFs 모형이 왜곡지표(VRIN, TRIN)를 비롯한 IRT, SVM 등의 모형 보다 다양한 기준점(cut-off value)에서 비정상 반응 패턴을 잘 구분할 수 있는지를 알아보기 위해 ROC curve에서의 AUC (Area Under the Curve)로 모형들의 성능을 비교하였다. 그 후 MRFs에서 비정상 반응 패턴을 판별할 수 있는 기준 값에 대하여 연구 하고, 그 기준 값 이하의 반응들을 비정상 반응 패턴으로 분류하였다.

연구 2는 MRFs를 활용한 응용연구이다. 대규모의 리서치 설문자료에서 MRFs를 활용하여 설문 길이에 따른 비정상 반응 패턴의 증가율을 알아보았다. 이를 위해 실제 산업장면에서 사용하고 있는 한국 리서치 설문 자료를 문항 수에 비례하여 이등분, 삼등분하여 분석하고, 좀 더 정밀한 분석을 위해 이동창(moving window)을 사용하여 설문 길이 혹은 문항 위치에 따라 비정상 반응 패턴이 증가함을 알아보하고자 한다.

1.2 논문의 구성

본 논문의 구성은 다음과 같다. 2장에서는 설문지에서의 비정상 반응 패턴을 탐지하는 기존 모형들을 살펴보고, 3장에서는 본 논문에서 제안하는 모형의 방법론적 근거와 사용된 MRFs 모형을 설명한다. 4장에서는 본 연구에서 모형을 평가하기 위해 사용한 정확도(ACC)와 ROC Curve에 대하여 설명한다. 5장에서는 연구 1에 대한 실험 소개 및 결과를 설명하고, 6장에서는 연구 2에 대한 실험 소개 및 결과를 설명한다. 마지막으로 7장에서는 연구 내용을 요약하고 향후 연구 과제와 함께 결론을 맺는다.

2. 선행연구

2.1 MMPI-2 타당도 척도

MMPI-2는 거짓 보고 및 무선(random) 반응 등을 탐지하기 위하여 여러 특수척도(타당도 척도)가 있다. 그 중 비정상 반응 패턴을 탐지하기 위해 1) 무선반응 비일관성 척도(Variable Response Inconsistency, VRIN)와 2) 고정반응 비일관성 척도(True Response Inconsistency, TRIN)를 사용한다.

무선반응 비일관성 척도(VRIN)

VRIN은 문항의 내용을 제대로 읽지도 않고 응답했거나 문항에 완전히 혹은 대부분 무작위로 응답한 사람들을 탐지하기 위한 척도이다(Butcher et al. 1989, 1995). 이 척도는 내용면에서 유사하거나 정반대인 문항들로 짝지어져 있으며, 비일관적으로 응답할 때마다 VRIN 척도의 원점수가 1점씩 높아진다. 총 67개의 문항반응 쌍들로 구성되어 있다.

고정반응 비일관성 척도(TRIN)

TRIN은 문항 내용과 상관없이 무분별하게 ‘그렇다’ 혹은 ‘아니다’로 응답하는 사람들을 탐지하기 위해 개발되었다(Butcher et al. 1989, 1995). VRIN 척도와 달리 서로 상반된 내용의 문항 쌍으로만 구성되어 있으며, 문항 쌍 내에서 상반된 내용임에도 불구하고 ‘그렇다’로 일관하거나 ‘아니다’로 일관할 경우 TRIN의 원점수는 1점씩 높아진다. 총 20개의 문항반응 쌍들로 구성되어 있다.

2.2 문항반응이론에 근거한 개인적합도

2.2.1. 문항반응이론

문항반응이론은 피험자의 특성(능력)과 문항의 특성을 동시에 고려하는 모형이다. 문항반응이론은 Rasch(1960)로부터 근원을 찾을 수 있다. 피험자의 특성(θ)과 문항의 여러 특성 중 난이도(difficulty)가 주어졌을 때, 반응 패턴(response pattern)을 로짓(logit)을 이용하여 확률로 나타낸 모형이다. Rasch 이후의 모형들에서 문항의 특성에 난이도 외에, 변별도(discriminating power), 추측도(pseudo chance level to guess) 등이 추가로 포함되었다. 편의상 난이도만 고려한 모형을 1PL(1-parameter logistic model), 난이도와 변별도를 고려한 모형을 2PL(2-parameter logistic model), 난이도와 변별도, 그리고 추측도를 고려한 모형을 3PL(3-parameter logistic model)이라 부른다.

설문지가 “그렇다(1)”와 “아니다(0)”의 이진(binary) 문항으로 주어졌을 경우, θ_j 의 특성을 가진 참여자가 각각 선택지에 반응할 확률은 식 (1)과 식 (2)와 같다.

$$\Pr(X_{ij} = \text{“그렇다(1)”} | \theta_j, \alpha_i, b_i, c_i) = c_i + (1 - c_i) \frac{1}{1 + \exp(-\alpha_i(\theta_j - b_i))} \quad (1)$$

$$\Pr(X_{ij} = \text{“아니다(0)”} | \theta_j, \alpha_i, b_i, c_i) = 1 - \Pr(X_{ij} = \text{“그렇다(1)”} | \theta_j, \alpha_i, b_i, c_i) \quad (2)$$

식 (1)에서 θ_j 는 j 번째 피험자의 능력을 나타내고, α_i , b_i , c_i 는 각각 i 번째 문항의 변별도, 난이도, 추측도를 나타낸다. 위의 수식은 3PL 모형으로서, 위 모형에서 추측도를 0으로 두면 2PL, 추측도와 변별도를 0으로 두면 1PL 모형이 된다. θ 는 보통 $-\infty < \theta < \infty$ 의 범위를 가지며,

평균이 0이고 표준편차가 1인 정규분포를 따르도록 정의한다. X_{ij} 는 베르누이 무선 변수(Bernoulli random variable)로서, $X_{ij} \in \{0,1\}$ 이다. 모형의 파라미터 추정에는 보통 EM 알고리즘을 활용한 MML(Marginal Maximum Likelihood)을 주로 사용한다.

본 연구에서 IRT를 사용한 연구 1의 경우 2PL 모형을 사용하였다. 1PL, 2PL, 3PL의 결정은 보통 우도비 검증(likelihood ratio test)으로 이루어지기도 하나 분석 이전에 고려되기도 한다. 추정해야할 파라미터 수가 많으면 더 많은 데이터를 요구하기 때문에 문항의 길이나 피험자 수에 따라 파라미터의 추정이 불안정할 수 있다. 본 연구 1의 경우 최대 약 450개의 반응 패턴이 있는데, 일반적으로 3PL의 경우 최소 1000명 이상의 사람들(sample size)을 필요하기 때문에(Gao and Chen, 2005; Hanson & Beguin, 2002) 2PL을 사용하여 분석하였다.

2.2.2. 문항반응이론의 가정 및 문제점

문항반응이론은 비정상 반응 패턴을 탐지하기 위해 만들어진 모형이 아니다. 즉, 모든 반응 패턴들이 정상적임을 전제로 하고 있다. 즉, IRT는 이 전제에 따라 다음 두 가지 가정을 가지고 있다. 1) 국지독립성(local independence) 가정과 2) 단조 증가성(monotonicity)의 가정이다.

국지독립성

식 (1) 과 식 (2) 를 식 (3) 처럼 간단히 나타낼 수 있고, j 번째 사람의 반응 패턴(X_j)에 대한 결합 확률(joint distribution)은 식 (4)로 나타낼 수 있다.

$$\begin{cases} P_i(\theta_j) & \text{if } X_{ij} = \text{“그렇다(1)”} \\ Q_i(\theta_j) & \text{if } X_{ij} = \text{“아니다(0)”} \end{cases} \quad (3)$$

$$\Pr(X_j = (X_{j1}, X_{j2}, \dots, X_{ji}) | \theta_j, a_i, b_i, c_i) = \prod_{i=1}^I P_i(\theta_j)^{X_{ij}} Q_i(\theta_j)^{(1-X_{ij})} \quad (4)$$

식 (4)는 ‘국지독립성(local independence)’의 가정을 나타낸다. 식 (4)는 특정 사람의 (연구자가 측정하고자 하는) 특성(능력)이 주어졌을 때, 서로 다른 문항에 반응할 확률은 독립이라는 것이다.

그러나, 비정상 반응 패턴은 (연구자가 측정하고자 하는) 특성(능력)과 무관하게 반응하고, 한 줄로 응답하기 등의 응답 패턴은 $\Pr(X_i | \theta_j)$ 와 $\Pr(X_{i+1} | \theta_j)$ 가 독립이 아니다. 따라서, 비정상 반응 패턴들이 많을 경우 문항반응이론의 최대우도 추정법을 사용한 파라미터 추정은 부정확 할 수 있다.

단조증가성

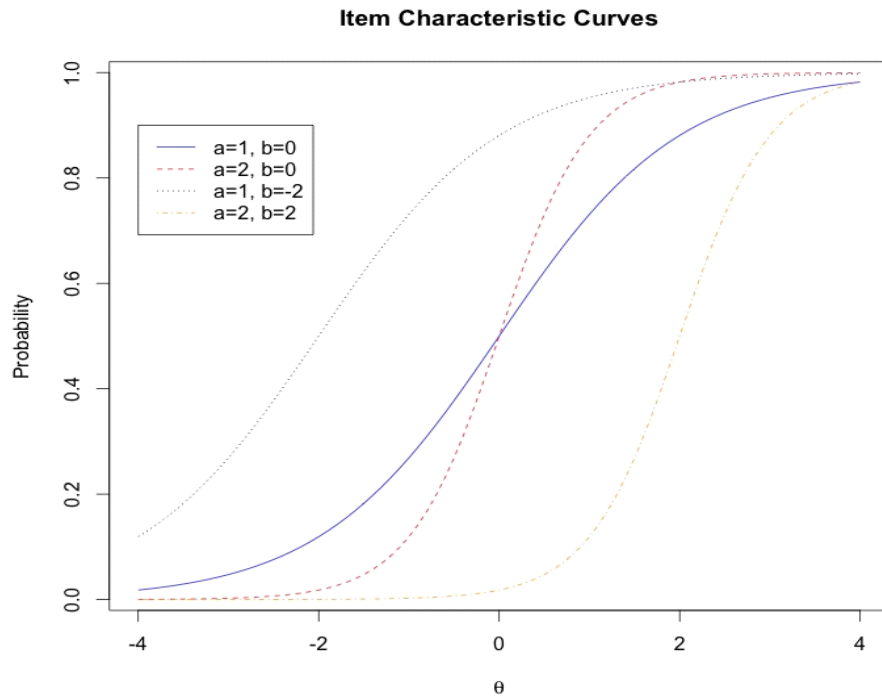


그림 1: 2모수 로지스틱 모형(2PL)의 Item Characteristic Curves
y축의 확률은 $\Pr(X_{ij} = \text{“그렇다(1)”} | \theta_j, a_i, b_i)$ 이다.

또한, 문항반응이론에서 사용하는 로지스틱 함수의 특성 상 단조증가성의 특징을 보인다. 즉, 그림 1 처럼 문항 변별도(a_i)가 0보다 크고, 문항 난이도(b_i)가 주어지면 특성(능력, θ)이 높은 사람이 특성(능력)이 낮은 사람에 비해 그 문항에서 “그렇다(1)”로 응답할 확률이 더 높아야 한다. 비정상 반응 패턴들은 (연구자가 측정하고자 하는) 특성과 관계없이 무작위로(randomly) 반응하기 때문에 특성(θ)이 높다고 해서 반드시 “그렇다(1)” 라고 반응할 확률이 높아지는 것은 아니다. 따라서, 이 가정 역시도 비정상 반응 패턴들이 많을 경우 모형의 파라미터를 불안정하게 추정하는 원인이 된다.

2.2.3. 문항반응이론을 이용한 개인적합도

문항반응이론에 근거한 개인적합도(Person-Fit Index)는 Levin & Rubin (1979)이 처음으로 제안하였다. 그들은 원래 특정 문항 패턴(X)에 대한 주변 확률(Marginal Probability)을 구하고자 하였다(식 5).

$$\Pr(X) = \int_{-\infty}^{\infty} f(X|\theta)g(\theta)d\theta \quad (5)$$

$g(\theta)$ 를 실제 설문지에 반응한 사람들에 대한 분포로 생각해보면, 수학적으로 최대우도법(Maximum Likelihood estimation)으로 추정된 $\hat{\theta}$, 즉 $\hat{\theta} = \operatorname{argmax}_{\theta} f(X|\theta)$ 를 중심으로 아주 작은 분산을 가진 $\tilde{g}(\theta)$ 로 대체할 수 있다. 즉, 특정 반응 패턴(X)은 특정 $\hat{\theta}$ 에서 관찰될 확률이 매우 높다고 할 수 있으며, 주변 확률 $\left(\int_{-\infty}^{\infty} f(X|\theta)\tilde{g}(\theta)d\theta \right)$ 은 조건부 확률 $(f(X|\hat{\theta}))^1$ 로 수렴할 수 있다(Levin & Rubin, 1979).

즉, 문항반응이론으로 추정된 개인의 특성($\hat{\theta}_j$)을 알면 그 사람의 반응 패턴에 대한 우도(Likelihood)를 구할 수 있다. 식 (6) 처럼 개인의 반응 패턴에 대한 우도에 로그를 취한 값을 문항반응이론에서의 Levin & Rubin의 개인적합도(l_0)로 정의한다(Levin & Rubin, 1979).

$$l_0(X) = \log f(X|\hat{\theta}) \quad (6)$$

$$l_0(X = (X_1, X_2, \dots, X_I) | \hat{\theta}) = \sum_{i=1}^I [X_i \log P_i(\hat{\theta}) + (1 - X_i) \log (1 - P_i(\hat{\theta}))] \quad (7)$$

1) $f(X|\hat{\theta})$ 는 식 (4)와 같이 정의한다.

식 (6)의 $f(X|\hat{\theta})$ 를 식 (4)를 이용하여 나타내면 식 (7)과 같다. 즉, Levin & Rubin의 개인적합도(l_0)는 특정 개인의 개별 문항에 대한 로그 우도를 전체 문항에 대하여 모두 합한 것이다.

Drasgow et al. (1985)는 추정된 개인의 특성($\hat{\theta}_j$)이 Levin & Rubin의 개인적합도(l_0)와 완전히 독립적이지 않을 수 있고, 개인별로 표준화된 개인적합도(l_z)가 표준화되지 않은 개인적합도(l_0) 보다 비정상 반응 패턴을 더 잘 탐지 할 수 있음을 밝혔다.

피험자의 반응이 이진(binary) 반응이라면, 특정 개인의 우도에 대한 평균과 분산은 식 (8) 과 식 (9) 와 같이 구할 수 있다.

$$E(l_0) = \sum_i [P_i(\hat{\theta})\log P_i(\hat{\theta}) + (1 - P_i(\hat{\theta}))\log(1 - P_i(\hat{\theta}))] \quad (8)$$

$$Var(l_0) = \sum_i [P_i(\hat{\theta})(1 - P_i(\hat{\theta}))\log[(P_i(\hat{\theta})/(1 - P_i(\hat{\theta})))^2] \quad (9)$$

이를 이용하여 표준화된 개인의 우도, 즉 $l_z = \frac{l_0 - E(l_0)}{[Var(l_0)]^{1/2}}$ 를 구할 수 있다. 보통의 경우 l_z 가 -2 보다 낮은 사람들을 비정상(aberrant) 반응 패턴으로 간주한다(Ferrando and Chico, 2001; Schmitt et al, 1999).

3. 마코프랜덤필드 모형

3.1. 마코프랜덤필드

마코프랜덤필드(Markov Random Fields, MRFs)는 문항반응이론과 달리 반응의 전이 역시 모형화 할 수 있다는 장점이 있다. 예를 들어, 문항의 특성과 관계없이 한 선택지로 일관된 반응('그렇다(1)' → '그렇다(1)' → '그렇다(1)')이나 문항의 특성과 관계없이 패턴이 있는 반응('그렇다(1)' → '아니다(0)' → '그렇다(1)' → '아니다(0)')의 경우 국지독립성에 위배된다.

MRFs은 흔히 알려진 마코프 체인(Markov chain), 은닉마코프모형(hidden markov model), 볼츠만 머신(Boltzmann machine), 이징 모형(Ising model), 홉필드 네트워크(Hopfield network)등을 일반화한 개념이다(Murphy. K, 2012). MRFs는 특히 컴퓨터 시각(computer vision) 분야에서 많이 사용되고 있다(Bishop, 2006; Murphy, 2012). MRFs의 변형인 조건부랜덤필드(Conditional Random Fields, CRF)를 포함하여, 1) 이미지 잡음제거(Bishop, 2006), 2) 이미지 내 영역 및 객체 구분(Zhao, 1992; Bhatt, 1994; Rother et al, 2004), 3) 영상 내 동작 구분(Quattoni, 2007) 등의 연구가 있다. 컴퓨터 시각 분야에서 MRFs의 연구가 활발한 이유는 1차원 구조를 나타내는 마코프 체인(Markov chain)과 달리, 2차원 혹은 3차원의 이미지 혹은 영상을 모형화 할 수 있기 때문이다.

그러나, MRFs 역시 마코프체인과 같이 1차원으로 정의할 수 있고²⁾ 텍스트와 같은 1차원 데이터에도 많은 연구들에서 응용되고 있다. 먼저, 자연언어처리의 경우 텍스트 자료가 있을 때 1) 특정 단어의 품사를 결

2) 하지만, 모든 마코프 체인은 1차원 MRFs이지만, 역이 항상 성립하는 것은 아니다(Chandgotia et al, 2014).

정(Lafferty et al, 2001), 2) 주제 분류(Zhu & Xing, 2010) 등에 사용되고 있다. 그 외 계산 생물학에서는 표준적인(standard) 염기 서열과 단백질 구조(sequence)를 추출하는데 사용되고 있다(Bernal et al, 2007).

MRFs 모형은 모두 마코프 속성(markov property)을 가지고 있다. 그림 2 와 같은 마코프 체인 모형에서의 마코프 속성은 미래의 상태는 현재의 상태에만 의존하고, 현재의 상태가 주어지면 과거와는 독립이다. 그러나, MRFs의 경우 이웃 구조에 의해서 영향을 받게 된다. 그림 3, 4, 5 는 각각 2차원 공간에서의 1차, 2차, N차 이웃 시스템을 나타낸다.

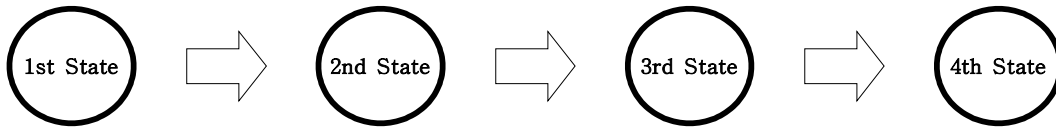


그림 2: 마코프 체인 모형

	v	
v	s	v
	v	

그림 3: 1차 이웃 시스템

v	v	v
v	s	v
v	v	v

그림 4: 2차 이웃 시스템

			6			
		5	4	3	4	5
		4	2	1	2	4
6	3	1	s	1	3	6
		4	2	1	2	4
		5	4	3	4	5
			6			

그림 5: N차 이웃 시스템(1,2,3,4,5,6)

3.2. 마코프랜덤필드의 가정 및 결합확률

MRFs는 변수 집합 $X = \{X_1, X_2, \dots, X_I\} \in \chi$ 의 결합분포를 나타내는 확률 그래프이다. 모든 무선 변수(random variable) X_i 는 특정 라벨 집합 L 에서의 f_i 값을 가지게 되고, $\Pr(X_1 = f_1, X_2 = f_2, \dots, X_I = f_I) = \Pr(X = f)$ 로 나타낼 수 있다. MRFs는 다음의 두 가지 속성을 만족해야 한다.

$$\Pr(X_i = f_i) > 0, \text{ all } i \in I \text{ (Positivity)} \quad (10)$$

$$\Pr(X_i = f_i | X_j = f_j, j \neq i, j \in I) = \Pr(X_i = f_i | X_j = f_j, j \in N_i) \quad (11)$$

(Markovianity = Locality)

식 (10)은 모든 사건(event)들의 확률이 0보다 커야하며, 식 (11)는 특정 사건($X_i = f_i$)을 제외한 모든 사건들이 주어졌을 때 특정 사건($X_i = f_i$)이 발생할 조건부 확률은 바로 특정 사건의 인접한 이웃 사건들(N_i)만 주어졌을 때 특정 사건($X_i = f_i$)이 발생할 조건부 확률과 같다. 즉, MRFs는 국소적 마코프 속성(Local Markov Property)을 가지고 있다.

MRFs는 무방향성으로 인해 연쇄 법칙(chain rule)을 사용하여, $\Pr(X_1 = f_1, X_2 = f_2, \dots, X_I = f_I) = \Pr(X = f)$ 의 결합 확률(Joint Probability)을 직접적으로 구할 수 없다. 그러나, MRFs는 Hammersley-Clifford theorem 등가성에 의해 깁스 분포(Gibbs distribution)와 동치가 되며, 깁스 분포(Gibbs distribution)는 식 (12)과 같이 정의된다.

MRFs에서의 무선변수를 노드(node)로 보고 이웃관계를 엣지(edge)로 볼 때 complete graph에서 무선 변수들의 집합을 클리크(clique)라고 정의한다. 따라서, 클리크는 모든 두 개의 노드들이 엣지로 연결되어 있는 subgraph이다. 이렇게 정의된 클리크의 포텐셜을 $V_c(f)$ 로 나타내며, 전체 에너지 함수는 $\sum_c V_c(f)$ 이다.

$$\Pr(X=f) = \frac{1}{Z} \exp(-\sum_c V_c(f)),$$

$$\text{where } Z = \sum_x \exp(-\sum_c V_c(f)) \quad (12)$$

또한, 이를 Factor Graph로 나타내면 식 (13) 과 같이 나타낼 수 있으며, 결합 확률인 $\Pr(X=f)$ 은 그래프에서의 maximal 클리크(A)의 포텐셜 함수(ψ_A)의 곱으로 간단히 나타낼 수 있다.

$$\Pr(X=f) = \frac{1}{Z} \prod_A \psi_A(f_A) \text{ where } Z = \sum_x \prod_A \psi_A(f_A), \psi_A(V_A) \geq 0 \quad (13)$$

따라서, MRFs에서 Hammersley-Clifford theorem 등가성에 의해 $\Pr(X_1=f_1, X_2=f_2, \dots, X_I=f_I) = \Pr(X=f)$ 결합 확률을 식 (13) 과 같이 구할 수 있다.

3.2. 이진(binary) 마코프랜덤필드 모형

3.2.1. 모형 상세화

본 연구에서는 이징 모형(Ising model)의 오류 함수를 변형한 이진(binary) 쌍(pair-wise) MRFs를 사용하였다. 이 모형은 전형적인(noise-free) 반응의 이행을 찾아내기 위해 고안되었지만, 이행을 찾아내기 위한 파라미터를 활용하여 개인적합도를 구할 수 있다.

모든 사람들의 반응 패턴들을 X 라고 하면, 그림 6 과 같이 나타낼 수 있다. 이때 j 는 사람, i 는 문항 번호이다. X 의 j 번째 행 벡터, 즉 $X_j = (x_{j1}, x_{j2}, \dots, x_{jI})$ 는 j 번째 사람의 반응 패턴이다.

$$X = \begin{bmatrix} x_{j_1 i_1} & x_{j_1 i_2} & \dots & x_{j_1 i_I} \\ x_{j_2 i_1} & x_{j_2 i_2} & \dots & x_{j_2 i_I} \\ \vdots & \vdots & \ddots & \vdots \\ x_{j_I i_1} & x_{j_I i_2} & \dots & x_{j_I i_I} \end{bmatrix}$$

그림 6: 모든 사람들의 반응 패턴

다음 그림 7은 문항들의 반응에 대한 모형이고, 확률은 식 (14)와 같다. 이 때 Z 는 정규화 상수(normalization constant)로서 모든 가능한 반응 패턴들을 반영하고 있다.

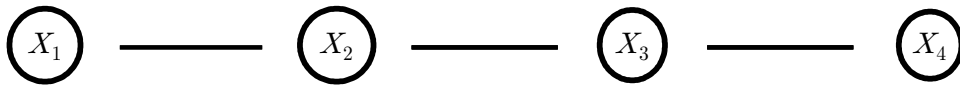


그림 7: 연구 모형 (문항들의 반응에 대한 모형)

$$\Pr(X_1, X_2, \dots, X_I) = \frac{1}{Z} \prod_i^I \psi_i(X_i) \prod_i^{I-1} \psi_{i,i+1}(X_i, X_{i+1}) \quad (14)$$

$$Z = \sum_{x_1} \sum_{x_2} \dots \sum_{x_I} \prod_{i=1}^I \psi_i(X_i) \prod_{i=1}^{I-1} \psi_{i,i+1}(X_i, X_{i+1})$$

보통 설문지를 만들 때, 설문지의 집중도 혹은 일관된 응답과 같은 편향을 감소시키기 위해 역문항(reverse item)을 사용한다. ‘반사회성’을 측정하기 위하여 다음과 같은 4 문항 설문지를 만들었다고 하자.

표 1. 설문지 구성 예시

문항번호(특징)	문항 내용
1번(정문항)	다른 사람들의 욕구나 감정에 둔감하다
2번(역문항)	호감이 느껴지고 괜찮은 인상을 준다
3번(정문항)	공격적으로 행동한다
4번(정문항)	자기중심적이다

반사회성을 가진 사람들은 1번, 3번, 4번 문항에서 “그렇다(1)”로 응답하도록 기대되며, 2번 문항에서는 “아니다(0)”로 응답하도록 기대된다. 따라서, 1번, 3번, 4번 문항은 정문항이라 정의하고, 2번 문항은 역문항이라 정의한다.

$$\psi_i(X_i = l) = (e^{\theta_i}, e^0)_{l+1} \quad (15)$$

3번, 4번 문항은 모두 ‘반사회성’을 측정하지만, ‘반사회성’을 측정하는 강도에 따라 $\psi_3(X_3 = \text{'그렇다(1)'}) \neq \psi_4(X_4 = \text{'그렇다(1)'})$ 일 수 있다. 즉, $\psi_i(x_i)$ 는 전체 사람들이 특정 문항의 보기를 선택할 확률을 반영한다. 따라서, 본 모형에서는 식 (15)의 e^{θ_i} 를 문항마다 모두 다르게 추정하였다.

이행 함수는 $\psi_{i,i+1}(X_i, X_{i+1})$ 또는 $\psi_{i,j}(X_i, X_j)$ ($j = i+1$)로 나타낸다. 이 때 2 종류의 $\lambda^{(k)}$ 를 사용되었다. 1) 정문항과 정문항 혹은 역문항과 역문항이 이어질 경우를 ‘정이행 관계’라고 정의하고, 2) 정문항과 역문항 혹은 역문항과 정문항이 이어질 경우를 ‘역이행 관계’라고 정의한다.

$\psi_{i,j}(X_i, X_j)$ 행렬에서, 1) ‘정이행 관계’의 경우에는 사람들이 응답을 바꿀 확률이 낮으므로 $\lambda^{(1)}$ 는 0 보다 작은 값을 가지게 되고, 반대로, 2) ‘역이행 관계’의 경우 사람들이 응답을 바꿀 확률이 바뀌지 않을 확률에 비해 높으므로 $\lambda^{(2)}$ 이 0 보다 큰 값을 가지게 된다.

$$\psi_{i,i+1}(X_i = l, X_{i+1} = m) = \begin{pmatrix} e^0 & e^{\lambda^{(k)}_{i,i+1}} \\ e^{\lambda^{(k)}_{i,i+1}} & e^0 \end{pmatrix}_{l+1, m+1} \quad (16)$$

$$\begin{cases} k=1 & \text{if } (i, i+1) - \text{정이행 관계} \\ k=2 & \text{if } (i, i+1) - \text{역이행 관계} \end{cases}$$

식 (15) 와 식 (16)의 파라미터들을 합쳐 $\Phi = [\Theta, \Lambda]$ 로 간단히 하면, 전체 사람에 대한 네거티브 로그 우도(negative log-likelihood)를 식 (17)로 간단히 나타낼 수 있다. 이 때 Z (partition function)는 문항이 적을 경우에는 직접 계산 가능하지만, 문항이 많을 경우에는 깁스 샘플러(single-site Gibbs sampler) 혹은 신뢰 전파 알고리즘(loopy belief propagation) 등을 사용하여 추론(inference)한다.

$$nll(\Phi) = \sum_{j=1}^J -\Phi^T X_j + \sum_{j=1}^J \log Z \quad (17)$$

3.2.2. 파라미터 추정법 및 개인적합도

파라미터 추정법

전체 우도($L(\theta)$)를 최대화하기 위한 파라미터 추정 방식은 (Liu and Nocedal, 1989)이 제안한 Quasi-Newton with Limited-Memory BFGS(Broyden - Fletcher - Goldfarb - Shanno)을 사용하였다. Newton 방식은 Hessian($d^2L(\Phi_n)$)을 직접구해야 하는데 데이터의 차원 (dimension)이 클 경우 Hessian의 계산 부담이 크다. Quasi-Newton 방식은 Hessian을 직접 구하지 않고 근사(approximation)하는 알고리즘이다. 즉, Hessian 대신 positive definite matrix(A_n)을 사용하여 파라미터의 값들을 업데이트를 한다.

먼저, 식 (19) 과 식 (20) 을 정의한다.

$$\nabla L(\Phi_n) - \nabla L(\Phi_{n+1}) \approx d^2L(\Phi_{n+1})(\Phi_n - \Phi_{n+1}) \quad (18)$$

$$s_n = \Phi_n - \Phi_{n+1} \quad (19)$$

$$g_n = \nabla L(\Phi_n) - \nabla L(\Phi_{n+1}) \quad (20)$$

BFGS에서는 secant 조건인 식 (21)을 만족하도록 A_{n+1} 을 식 (22)에 의해 업데이트 한다(Lange, 2010). 본 연구에서 A_n 의 초기값을 positive definite matrix인 $A_0 = I$ 로 주었다.

$$-A_{n+1}s_n = g_n \quad (21)$$

$$A_{n+1} = A_n + \frac{g_n g_n^T}{g_n^T s_n} - \frac{A_n s_n s_n^T A_n}{s_n^T A_n s_n} \quad (22)$$

제한된 메모리(Limited-Memory) 방식은 A_{n+1} 을 구할 때 A_n 을 모두

사용하는 것이 아니라, 메모리에 저장해 둘 C (Correction Parameter)를 두어 가장 최근의 s_n 과 g_n 을 C 개수 만큼만을 이용하여 업데이트를 할 수 있다. 즉, 계산 복잡도를 $O(n^2)$ 에서 $O(C \times n)$ 으로 줄일 수 있다 (Nocedal and Wright, 1999).

MRFs를 이용한 개인 적합도

추정된 (θ, λ) 의 파라미터를 이용하여, j 번째 사람의 반응 패턴에 대하여 개인별 우도를 계산할 수 있다. 만약, j 번째 사람이 비정상적이라면 정상적으로 응답한 사람들에 비해 $\psi_i(X_i)$ 와 $\psi_{i,j}(X_i, X_{i+1})$ 에서 낮은 파라미터 값에 해당하는 반응을 선택할 가능성이 높다. 따라서, 비정상 반응 패턴을 보인 사람의 개인별 우도는 낮을 것이다.

좀 더 명확한 이해를 돕기 위하여, 다음과 같은 예시를 들어보자.

먼저 노드 포텐셜(ψ_i)을 표 2와 같이 구하였다고 해보자. 또한, $\psi_1(X_1 = \text{"그렇다(1)}) = .4$, $\psi_1(X_1 = \text{"아니다(0)}) = .6$ 라고 할 때, 계산 편의를 위해 상수비 2 : 3 으로 치환해보자.

표 2: ψ_i (노드 포텐셜)

문항 번호	그렇다(1) (확률)	아니다(0) (확률)	그렇다(1) (상수비)	아니다(0) (상수비)
1번	.4	.6	2	3
2번	.7	.3	7	3
3번	.2	.8	1	4
4번	.4	.6	2	3

그 다음, 엣지 포텐셜($\psi_{i,i+1}$)을 표 3 과 표 4와 같이 구하였다고 해보자.

‘정 이행 관계’

표 3: ‘정 이행 관계’의 $\psi_{i,i+1}$ (엣지 포텐셜)

전이	그렇다(1) (확률)	아니다(0) (확률)	그렇다(1) (상수비)	아니다(0) (상수비)
그렇다(1)	.7	.3	7	3
아니다(0)	.3	.7	3	7

‘역 이행 관계’

표 4: ‘역 이행 관계’의 $\psi_{i,i+1}$ (엣지 포텐셜)

전이	그렇다(1) (확률)	아니다(0) (확률)	그렇다(1) (상수비)	아니다(0) (상수비)
그렇다(1)	.3	.7	3	7
아니다(0)	.7	.3	7	3

이 때, 각각 반응 패턴에 대하여 포텐셜의 곱을 구해보면 표 5와 같다. 표 5에서는 “그렇다(1)”를 “+”, “아니다(0)”를 “-”로 나타내었다.

표 5: 모든 반응 패턴들에 대한 포텐셜 값

문항번호				ψ_i				$\psi_{i,j}$			포텐셜 곱
(1)	(2)	(3)	(4)	np (1)	np (2)	np (3)	np (4)	ep (1)	ep (2)	ep (3)	
+	+	+	+	2	7	1	2	3	3	7	1764
-	+	+	+	3	7	1	2	7	3	7	6174
+	-	+	+	2	3	1	2	7	7	7	4116
-	-	+	+	3	3	1	2	3	7	7	2646
+	+	-	+	2	7	4	2	3	7	3	7056
-	+	-	+	3	7	4	2	7	7	3	24696
+	-	-	+	2	3	4	2	7	3	3	3024
-	-	-	+	3	3	4	2	3	3	3	1944
+	+	+	-	2	7	1	3	3	3	3	1134
-	+	+	-	3	7	1	3	7	3	3	3969
+	-	+	-	2	3	1	3	7	7	3	2646
-	-	+	-	3	3	1	3	3	7	3	1701
+	+	-	-	2	7	4	3	3	7	7	24696
-	+	-	-	3	7	4	3	7	7	7	86436
+	-	-	-	2	3	4	3	7	3	7	10584
-	-	-	-	3	3	4	3	3	3	7	6804
합계											Z=18390

각 반응 패턴에 대한 확률은 포텐셜의 곱을 Z(포텐셜 곱들의 합)로 나누어 구할 수 있다. 각 반응 패턴에 대한 확률을 내림차순으로 정리하면 표 6과 같다.

표 6: 모든 반응 패턴들에 대한 확률

(1)	(2)	(3)	(4)	Pr	순위
-	+	-	-	.456	1
-	+	-	+	.130	2
+	+	-	-	.130	2
+	-	-	-	.056	4
+	+	-	+	.037	5
-	-	-	-	.036	6
-	+	+	+	.032	7
+	-	+	+	.022	8
-	+	+	-	.021	9
+	-	-	+	.016	10
-	-	+	+	.014	11
+	-	+	-	.014	11
-	-	-	+	.010	13
+	+	+	+	.009	14
-	-	+	-	.009	15
+	+	+	-	.006	16
합계				1	

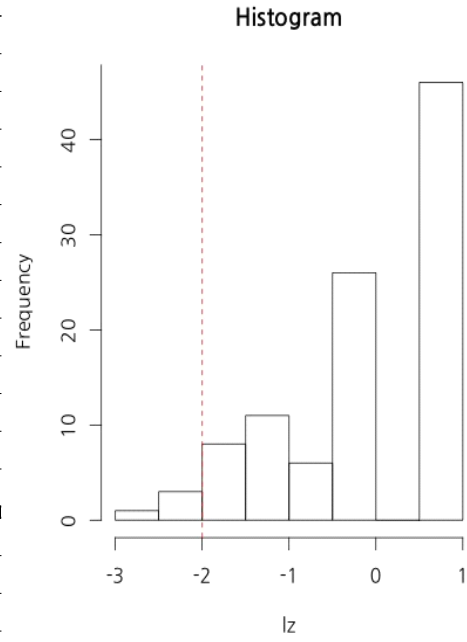


그림 8: MRFs 개인적합도 히스토그램

J 명의 반응 패턴들이 각 반응 패턴 확률에 따라 존재할 때, 그림 8은 100명의 개인별 로그 우도를 표준화한 히스토그램이다. 표준화된 개인별 로그 우도(l_z)를 개인적합도라 정의하고, 이 값 중 특정 값, 예를 들어, -2 보다 낮은 반응 패턴을 가려내면 13 ~ 16순위의 비정상 반응 패턴을 찾을 수 있다. -2 등의 기준에 대해서는 5.2.2에서 상세히 논의할 것이다.

4. 모형 평가 지표

모형의 성능을 평가하기 위한 여러 지표들 중 본 연구에서는 Receiver Operating Characteristics(ROC) curve에서의 곡선하면적(Area Under the Curve, AUC)와 ACC(Accuracy) 및 B. ACC(Balanced Accuracy)를 사용하여 모형의 성능을 평가하였다.

먼저, 특정 모형에서 이진(binary)으로 예측한 라벨과 실제 라벨의 관계를 2×2 매트릭스 형태로 나타내면 표 7과 같다.

표 7: Confusion Matrix 예시

		실제 라벨	
		비정상	정상
모형이 예측한 라벨	비정상	True Positive	False Positive
	정상	False Negative	True Negative

(1) True Positive(TP)는 실제로 ‘비정상 반응 패턴’ 중에서 모형이 예측한 ‘비정상 반응 패턴’의 수이다.

(2) True Negative(TN)은 실제로 ‘정상적인 반응 패턴’ 중에서 모형이 예측한 ‘정상적인 반응 패턴’의 수이다.

(3) False Positive(FP)은 실제로 ‘정상적인 반응 패턴’임에도 불구하고 모형이 ‘비정상 반응 패턴’으로 예측한 수이다. 통계학적으로 ‘Type I error’에 해당된다.

(4) False Negative(FN)은 실제로 ‘비정상 반응 패턴’임에도 불구하고 모형이 ‘정상적인 반응 패턴’으로 예측한 수이다. 통계학적으로 ‘Type II error’에 해당된다.

분류 모형(classification model) 또는 분류기(classifier)의 정확도(Accuracy)는 식 (23)과 같이 정의된다.

$$\text{정확도(ACC)} = \frac{TP + TN}{allP + allN} \quad (23)$$

두 클래스(class)의 크기가 서로 상이할 경우 정확도(ACC)를 사용하여 모형의 성능을 평가하면 편향이 발생할 수 있다(Daskalaki., et al, 2006; Elazmeh., et al, 2006). 따라서, 추가적으로 Balanced ACC(B. ACC)를 사용하여 두 모형의 정확도를 탐지하였다. B. ACC는 식 (24)와 같이 정의된다.

$$\text{Balanced ACC(B. ACC)} = \frac{TPR + TNR}{2} \quad (24)$$

(5) True Positive Rate(TPR 또는 sensitivity)은 $TP / (TP + FN)$ 을 나타낸다.

(6) True Negative Rate(TNR 또는 specificity)는 $TN / (FP + TN)$ 을 나타낸다.

ROC Curve 와 AUC

IRT의 개인적합도(반응 패턴의 로그 우도)와 MRFs의 개인적합도(반응 패턴의 로그 우도)는 비정상 반응 패턴으로 분류하는 컷오프 기준점(threshold)에 따라 정확도가 달라질 수 있다. Schmitt et al. (1999)은 IRT의 컷오프 기준인 -2가 1) 비정상 반응 패턴의 비율 2) 설문지의 길이 3) 문항 특성 등에 따라 비일관적임을 밝혔다. MRFs 역시 IRT와 마찬가지로 본 연구에서는 제안하는 컷오프 기준이 정확하지 않을 수 있다. 따라서, 모든 컷오프의 기준(threshold)을 반영하고 있는 ROC curve를 통해 모형의 변별도를 평가하는 것이 선행되어야 한다.

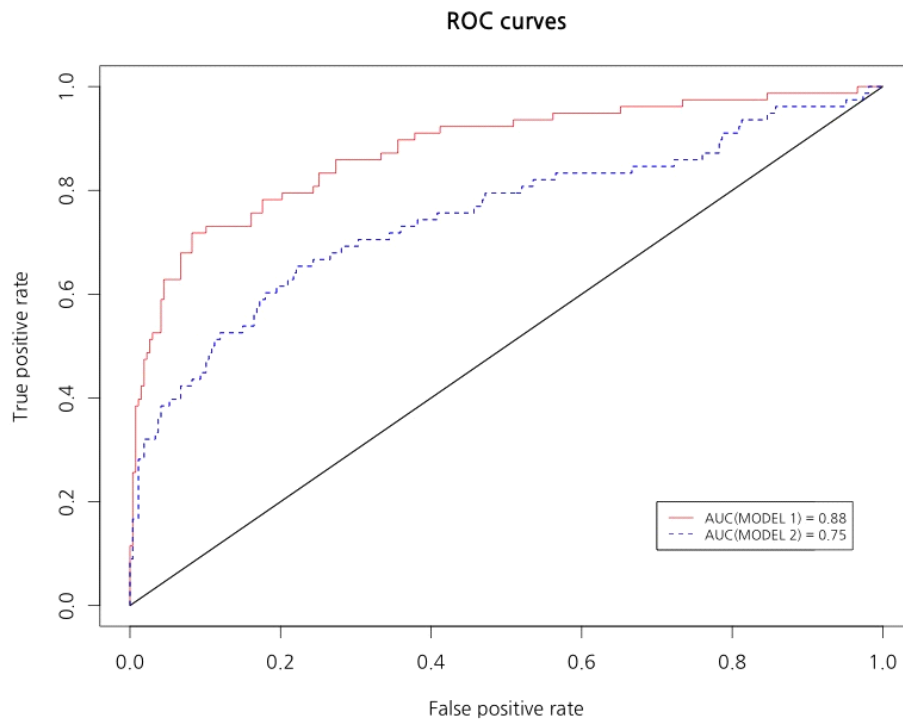


그림 9: ROC curve 예제

ROC(Receiver Operating Characteristics, ROC) curve는 어떤 모형의 이진 판단결과(binary classifier)의 성능(performance)를 보여주는 그래프이다(Fawcett, 2006). 기계 학습(machine learning) 분야에서는 일찍이

Spackman (1989)이 알고리즘들을 평가하기 위해 ROC curve를 사용하였고, 최근 기계학습 커뮤니티에서 ROC curve의 사용이 증가하고 있다 (Frawcett, 2006).

ROC curve는 X축을 False Positive Rate, Y축을 True Positive Rate로 나타낸다. 즉, 컷오프 값의 변화에 따른 False Positive Rate에 따라 True Positive Rate가 어떻게 달라지는 지를 보여준다. ROC curve의 곡선하면적(Area Under the Curve)을 구하면 어떤 모형이 더 변별도가 높은지를 알 수 있다. 정확도(Accuracy, ACC)의 경우 종종 좋지 못한 (poor) 성능을 보일 수 있으며(Provost and Fawcett, 1997; Provost et al., 1998), AUC는 식 (23)의 ACC보다 통계적으로 더 일관되고 더 변별력 있는 지수이다(Ling et al., 2003).

AUC는 분류기(classifier)를 비교하는 단일 스칼라 지수이다. 통계학적으로, AUC는 분류기가 무작위(randomly)로 negative instance(정상 반응패턴)를 고르는 것 보다 더 무작위로 positive instance(비정상 반응패턴)를 고를 것에 대한 확률을 나타낸다. 본 연구에서는 비정상 반응 패턴을 탐지하는 것이기 때문에 비정상 반응 패턴이 positive instance가 된다. AUC는 확률이므로 0과 1사이에 존재하고, AUC가 1에 가까울수록 더 좋은 분류기, 즉 분류를 잘할 수 있는 모형이라고 하겠다.

Hand and Till(2001)은 Wilcoxon-Mann-Whitney (WMW) two-sample 검증 통계치와 동일함을 밝혔고 AUC는 식 (25)와 같다.

$$AUC = \frac{S_+ - p(p+1)/2}{pn} \quad (25)$$

식 (24)에서 p 와 n 은 각각 positive(비정상) 와 negative(정상) 반응패턴을 각각 말하며, S_+ 은 모형(classifier)의 값을 오름차순 정렬했을 때, positive(비정상 반응 패턴)의 랭크(rank)의 합을 의미한다.

5. 연구 1

본 연구 1에서는 1) MRFs 모형이 비정상 반응 패턴을 잘 변별할 수 있는지를 알아보기 위해 여러 다른 모형과 ROC curve의 AUC를 비교하여 평가한다. AUC 분석에서는 training set과 test set을 각각 2/3과 1/3으로 나누어 분석하였다. 2) MRFs 모형이 잘 변별할 수 있는 컷-오프 기준을 찾고 모형의 정확도(ACC 및 B. ACC)를 평가한다. ACC 및 B. ACC 분석에서는 training set과 test set을 나누지 않고 분석하였다.

SVM의 경우 1) AUC를 활용한 분석에서는 two-class 방식을 2) ACC 및 B. ACC 분석에서는 one-class 방식을 사용하였다.

5.1. 설문자료 및 실험방법

설문자료

세계적으로 가장 널리 쓰이고 가장 많이 연구되어 있는 검사로서, MMPI-2의 일차 목적은 자기보고식 설문지를 통해 정신과적 진단분류를 하는 것이다. 한국판 MMPI-2는 (주)마음사랑에서 2004년도부터 표준화 작업이 본격적으로 시작되어 2005년도에 정식 출판되었다.

실험방법

연구 1에서는 장은경(2010)이 수집한 자료를 재분석하였다. 이 자료는 정상 반응 패턴들과 비정상 반응패턴들이 존재한다. 실험방법은 다음과 같다.

먼저, 서울대학교 학부생 225명을 대상으로 한국판 MMPI-2 설문지를 주고 정상적으로 설문지에 반응하게 하여 정상 반응 패턴들을 모은다.

1주일 뒤에 동일한 사람을 대상으로 설문지 없이 답안지만 제시하고 마음대로 응답을 하도록 지시 하였다. 정상적으로 설문지에 반응한 반응 패턴들을 정상 반응 패턴으로 정의하고, 마음대로 응답한 반응 패턴들을 비정상 반응 패턴으로 정의한다.

분석에 사용한 코드

모든 분석은 통계 프로그램인 R(ver. 3.1.1)을 사용하였다. IRT는 ltm 라이브러리를 사용하였다. 또한, two-class SVM은 e1071, one-class SVM은 kernlab 라이브러리를 사용하였다. MRFs는 Schmidt(2012)의 UGM(Undirected Graphical Models) 매트랩(matlab) 코드를 Wu(2014)가 R로 변형한 CRF(conditional random fields) 라이브러리를 사용하였다. 단, 파라미터 업데이트 방식은 ‘Limited Memory - BFGS’ 방식을 사용하였다.

5.2. 실험결과

5.2.1. AUC를 이용한 모형의 변별도 평가

비정상 반응 패턴(총 225개)과 정상 반응 패턴(총 225개)에 대하여 왜곡지표(VRIN, TRIN의 점수), IRT, MRF의 표준화된 로그 우도(l_z)에 대하여 ROC curve의 AUC를 구하였다.

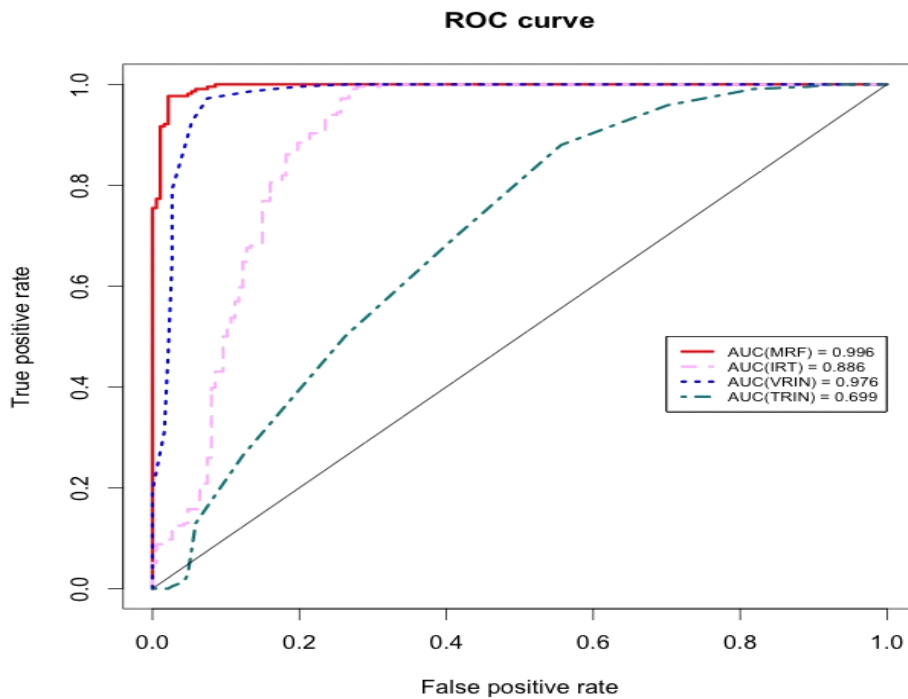


그림 10: VRIN, TRIN, MRFs, IRT 모형 민감도 비교

그림 10은 VRIN, TRIN, MRFs, IRT 모형의 AUC이며, AUC는 MRF(.996), VRIN(.976), IRT(.886), TRIN(.699)의 순으로 나타났다. 또한 모든 컷-오프 값에서 MRFs 모형이 다른 모형에 비해 더 비정상 반응 패턴을 잘 탐지하였다.

세부 연구1. 비정상 패턴 비율 조정

대부분 설문지의 경우 비정상적으로 응답하는 사람들의 비율이 50% 미만이므로, 비정상패턴비율 = (비정상 응답 패턴 수) / (전체 응답 패턴 수)을 25%, 20%, 15%, 10%, 5%로 변경가면서 모형 민감성을 다시 비교하였다. 또한, 일반적인 설문지의 경우 왜곡지표(VRIN, TRIN)가 없기 때문에 왜곡지표를 제외한 총 456문항에 대해서도 분석하였다. 또한, 엄밀한 비교를 위해 training set을 2/3로 test set을 1/3로 나누어 재분석하였다

세부 연구(1 ~ 3)에서는 감독학습이지만 대부분의 변별에서 정확도가 높은 지지벡터머신(SVM)도 함께 분석하였다. AUC를 이용한 모형 분석에서 SVM은 가우시안 커널을 사용하였고, two-class 방식을 사용하였다. SVM의 variance와 cost는 1로 고정하였다.

표 8: VRIN, TRIN, IRT, MRFs, SVM의 평균 AUC(총 567문항)
(괄호 안은 표준편차, 총 10회 시행)

비정상 반응 비율	VRIN	TRIN	왜곡 지표 유무 포함	IRT	MRFs	SVM 2 Class
25%	.976 (.02)	.699 (.06)	(567문항)	.861 (.012)	.996 (.002)	.998 (.001)
			제거 (456문항)	.819 (.011)	.996 (.002)	.999 (.002)
			포함 (567문항)	.901 (.005)	.998 (.001)	1 (0)
20%	.973 (.05)	.663 (.08)	제거 (456문항)	.858 (.013)	.996 (.003)	.998 (.002)
			포함 (567문항)	.961 (.003)	.998 (.002)	.999 (.002)
			제거 (456문항)	.944 (.003)	1 (0)	.999 (.002)
15%	.982 (.02)	.671 (.25)	포함 (567문항)	.974 (.011)	.998	1 (0)
			제거 (456문항)	.901 (.012)	1 (0)	.999
			포함 (567문항)	.994 (.010)	1 (0)	1 (0)
10%	.973 (.06)	.692 (.15)	제거 (456문항)	1 (0)	1 (0)	1 (0)
			포함 (567문항)	.994 (.010)	1 (0)	1 (0)
			제거 (456문항)	1 (0)	1 (0)	1 (0)
5%	.956 (.09)	.786 (.19)	포함 (567문항)	.994 (.010)	1 (0)	1 (0)
			제거 (456문항)	1 (0)	1 (0)	1 (0)
			포함 (567문항)	.994 (.010)	1 (0)	1 (0)

표 8 을 보면, (1) IRT의 경우 비정상 반응 패턴의 비율이 많아질수록 AUC가 감소하고 있다. (2) MRFs와 SVM은 설문지에서 측정할 수 있는 왜곡 지표(VRIN, TRIN) 보다 더 우수하였다. (3) 왜곡 지표(VRIN, TRIN)를 제거하였을 경우, MRFs와 SVM의 AUC는 감소하지 않으나, IRT의 경우 AUC가 감소하였다. (4) 567문항을 모두 분석했을 경우 비정상 비율이 10% 미만이면 IRT, MRFs, SVM 3개의 모형 모두 거의 100%의 정확도를 보였다.

세부 연구2. 적은 문항(총 30문항)

그러나, 문항이 적더라도 효과적으로 무선반응을 탐지할 수 있어야 하므로, VRIN과 TRIN을 제외한 456문항 중에서 문항의 수를 30개로 줄여서 재분석하였다. 연속적인 30개의 문항을 총 10회의 시행마다 서로 다른 set을 추출하였다.

표 9: IRT, MRFs, SVM의 평균 AUC (총 30문항)
(괄호안은 표준편차, 총 10회 시행)

비정상 패턴 비율	AUC		
	IRT	MRFs	SVM 2-class
9%	.912	.985	.991
	(.05)	(.02)	(.01)
7%	.961	.996	.987
	(.05)	(.02)	(.02)
5%	.954	.994	.993
	(.03)	(.02)	(.02)
3%	.966	.989	.987
	(.04)	(.02)	(.02)

표 9 는 비정상 패턴의 비율이 10%미만에 대하여 분석한 결과이다. 표 9 를 보면, 30문항의 적은 문항에서도 MRFs와 SVM의 성능이 IRT 보다 더 좋았다.

세부 연구3. 부분 비정상 반응 패턴에 대한 연구(총 30문항)

모형의 성능을 좀 더 정밀하게 평가하기 위해 '정상 반응 패턴들'만을 가지고 모의 실험을 통해 세 모형을 비교하였다. 문항 전체에 대하여 비정상적으로 반응하는 것뿐만 아니라 일부 문항만을 비정상적으로 반응할 경우에도 비정상 반응 패턴을 잘 탐지하는지를 알아보기 위한 연구이다.

모의 실험은 총 2종류이다. 모의 실험 1은 사람이 설문지에 반응을 하다가 중간 중간 무작위로 반응하도록 조작하였고, 모의 실험 2는 연속적인 문항을 한 번호로 반응하도록 조작하였다.

모의 실험 1. 무선 문항 및 무선 반응 조작

모의 실험 1은 3단계로 이루어진다. 왜곡 지표를 제외한 456문항 중에서 1) 무작위로 연속 30문항을 샘플링한 후, 2) 정상반응들 중에서 7%, 5%, 3%의 반응패턴들을 각각 샘플링하였다. 3) 456문항 중 30%, 50%, 70%의 문항들을 샘플링하여 개인 그리고 문항마다 독립적으로 1/2의 확률로 “그렇다(2)” 혹은 “아니다(1)”로 개인의 응답을 조작하였다. 표 10에서 ‘+’ 와 ‘-’는 ‘그렇다’ 와 ‘아니다’의 조작된 반응을 의미한다. 즉, 조작 조건은 비정상패턴비율(3조건) \times 문항조작비율(3조건)이다. 1), 2), 3)의 과정을 총 1000회 반복하였다. 이 때, 비정상패턴비율 = r / N , 문항조작비율 = e / I 이다.

표 10: 모의 실험 1. 무선 문항 및 무선 반응 조작. ‘+’와 ‘-’는 조작된 문항을 말한다. 문항의 순서는 무선(random)적으로 선별되었으며, 문항의 조작 역시 개인과 문항에 대하여 1/2의 확률로 무작위로 조작하였다.

전체 사람 수 (N)	전체 문항수(I)								정상
		+			+		-		조작 (r)
			+			-		+	
		+			+		-		
	+	-	-						
	(e)								

모의 실험 2. 연속 문항 및 연속 반응 조작

모의 실험 2 역시 3단계로 이루어진다. 왜곡 지표를 제외한 456문항 중에서 모의 실험 1과 마찬가지로 3단계에 걸쳐 반응을 조작한다. 1) 과 2)의 과정은 모의 실험 1과 동일하며, 3)은 456문항 중 30%, 50%, 70%의 연속적인 문항들을 샘플링하여, 패턴마다 1/2의 확률로 연속적으로 “그렇다(2)” 혹은 “아니다(1)”로 개인의 응답을 조작하였다. 조작 조건은 비정상패턴비율(3조건) \times 문항조작비율(3조건)이다. 1), 2), 3)의 과정을 총 1000회 반복하였다. 이를 도식화하면 표 11과 같다. 이 때, 비정상패턴비율 = r / N , 문항조작비율 = e / I 이다.

표 11: 모의 실험 2. 연속 문항 및 연속 반응 조작. ‘+’와 ‘-’는 조작된 문항을 말한다. 문항의 순서는 연속적으로 선별되었으며, 문항의 조작 역시 한 개인은 동일한 반응을 반복한다. 개인의 연속적인 반응에 대하여 1/2의 확률로 무작위로 조작하였다.

		전체 문항수(I)							
전체 사람 수 (N)									정상
		+	+	+					조작 (r)
			+	+	+				
						-	-	-	
	-	-	-						
	(e)								

분석 결과, 그림 11을 보면 모의실험 1의 경우 SVM 모형이 IRT와 MRFs 비해 전반적으로 변별도가 근소하게 낮고, 그림 12를 보면 모의 실험 2의 경우 IRT 모형이 MRFs와 SVM 모형보다 변별도가 낮으며, 특히 문항조작비율이 더 높을수록, 즉 한 줄 연속 반응이 길수록 IRT 모형의 판별 정확도가 낮아짐을 알 수 있다.

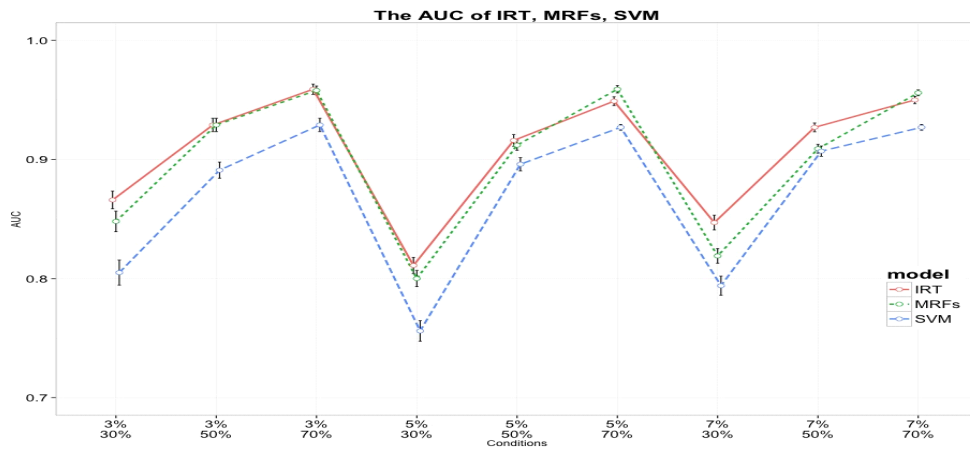


그림 11: 모의 실험 1(무선 문항 및 무선 반응 조작) 결과. IRT, MRFs, SVM 모형의 AUC 비교, x축의 조건은 비정상사람비율(3%, 5%, 7%) × 문항조작비율(30%, 50%, 70%)이다.

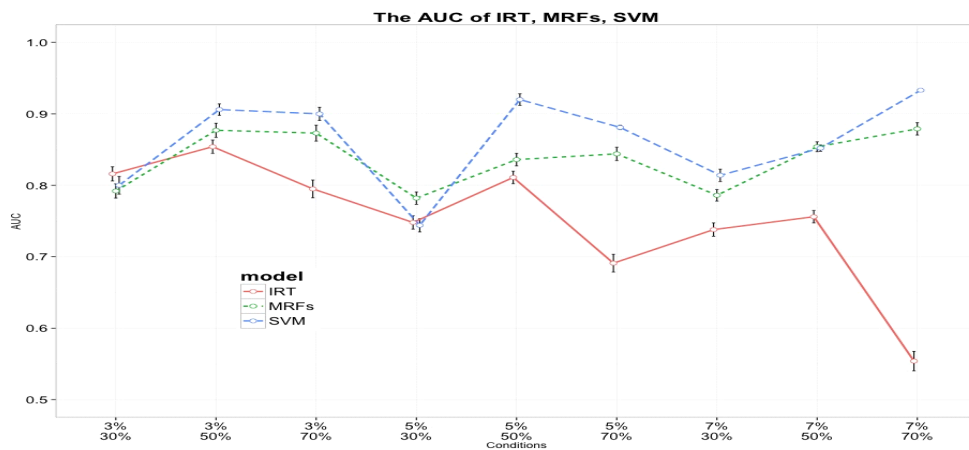


그림 12: 모의 실험 2(연속 문항 및 연속 반응 조작) 결과. IRT, MRFs, SVM 모형의 AUC 비교, x축의 조건은 비정상사람비율(3%, 5%, 7%) × 문항조작비율(30%, 50%, 70%)이다.

논의

5.2.1.에서는 AUC를 이용하여 MRFs 모형이 다른 모형에 비해 비정상 반응 패턴들을 민감하게 구분할 수 있는지를 평가하였다. 연구 결과, 여러 조건들에서 MRFs 모형이 일관되게 비정상 반응 패턴들을 잘 변별할 수 있음을 검증하였다.

그러나, 실제 연구 장면에서 MRFs 모형으로 비정상 반응 패턴들을 판별하기 위해서는 비정상 반응 패턴과 정상 반응 패턴을 잘 판별할 수 있는 MRFs 개인적합도의 컷오프 기준(값)을 구할 필요가 있다. 따라서, 5.2.2.에서는 MRFs 모형에서의 컷오프 기준에 대하여 연구하였다.

5.2.2. 마코프랜덤필드 모형에서의 컷오프 기준

본 연구에서는 MRFs 모형에서의 컷오프 기준을 두 가지로 제시하고 있다. 첫째, 반응 패턴에 대한 기본(base) 모형을 생각해볼 수 있다. 모든 문항에서의 각 반응(1 또는 0)에 대한 확률이 모두 iid하다면, 그 반응 패턴의 로그 우도는 $\log(.5) \times I$ (I = 문항수)가 될 것이다. MRFs 모형은 각 반응 패턴의 우도를 구할 수 있기 때문에, 기본 iid 모형에서 모든 문항에 대하여 무선적으로 반응한 반응 패턴의 우도와 비교할 수 있다. 즉, $\log(.5) \times I$ 이하의 값들을 비정상 반응 패턴으로 분류할 수 있다.

그러나, 모든 문항이 아닌 부분적으로 비정상 반응 패턴을 보이는 경우 우의 기준($\log(.5) \times I$)이 정확하지 않을 수 있다. 또한, 설문지 특성에 따라 기본 모형의 기준은 다르게 적용될 수 있다. 따라서, 모형의 판별 정확도가 높은 컷오프 지점을 탐지하기 위해 컷오프 기준(threshold)에 따른 정확도(ACC)를 세부 연구(1 ~ 3)에 걸쳐서 연구하였다. 단, 비정상 패턴 비율에 따라 컷오프 기준이 달라질 수 있기 때문에 비정상 패턴 비율이 10% 미만인 경우에 한정해서 세부 연구를 하였다.

세부 연구1. 비정상 패턴 비율 조정

먼저 총 567문항으로, 비정상사람비율을 3%, 5%, 7%, 9%로 달리하였을 때, 컷오프 기준에 따른 ACC는 그림 13 과 같다. 4 조건을 동시에 고려했을 때 최적의 컷오프 기준은 -2 부근으로 나타난다.

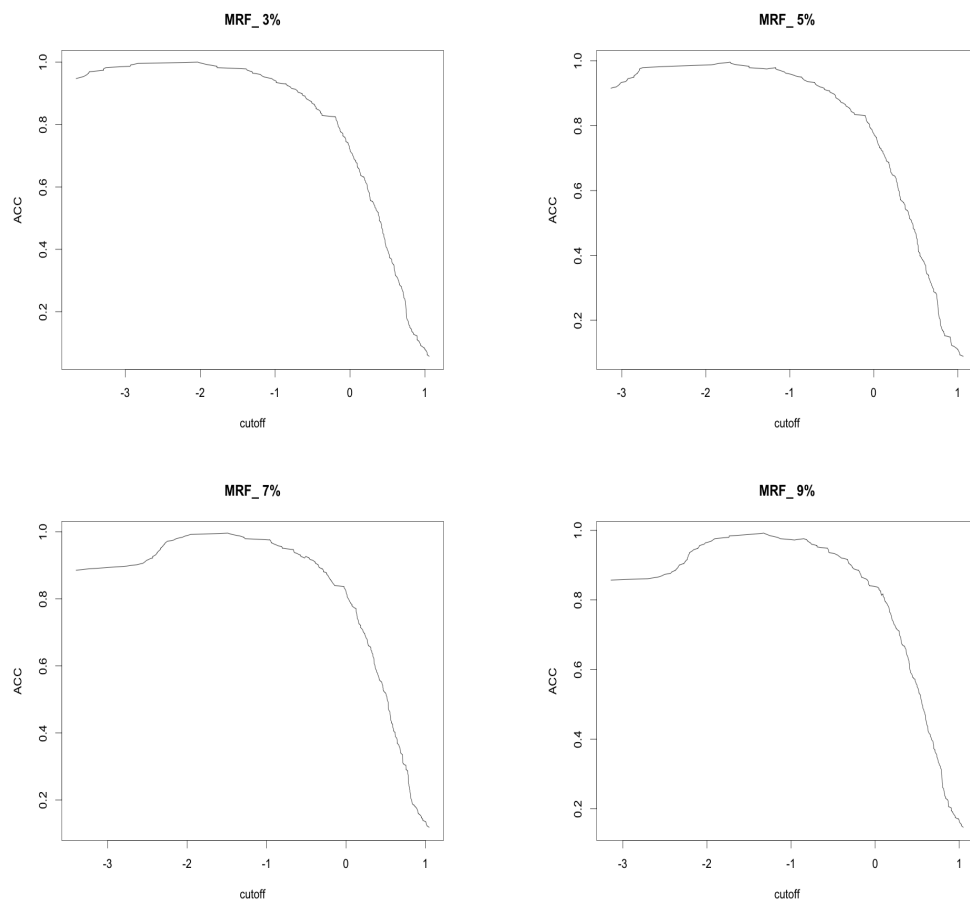


그림 13: 총 567문항에 대한 컷오프 기준에 따른 ACC

세부 연구2. 적은문항(30문항)

왜곡지표를 제외한 456문항 중에서, 비정상사람비율을 3%, 5%, 7%, 9%로 달리하면서 연속적인 30문항을 샘플링하여 총 100회 반복하였다. 그림 14는 ACC의 평균이며, 여러 조건을 동시에 고려했을 때 최적의 컷오프 기준은 -2 ~ -1 부근으로 나타난다.

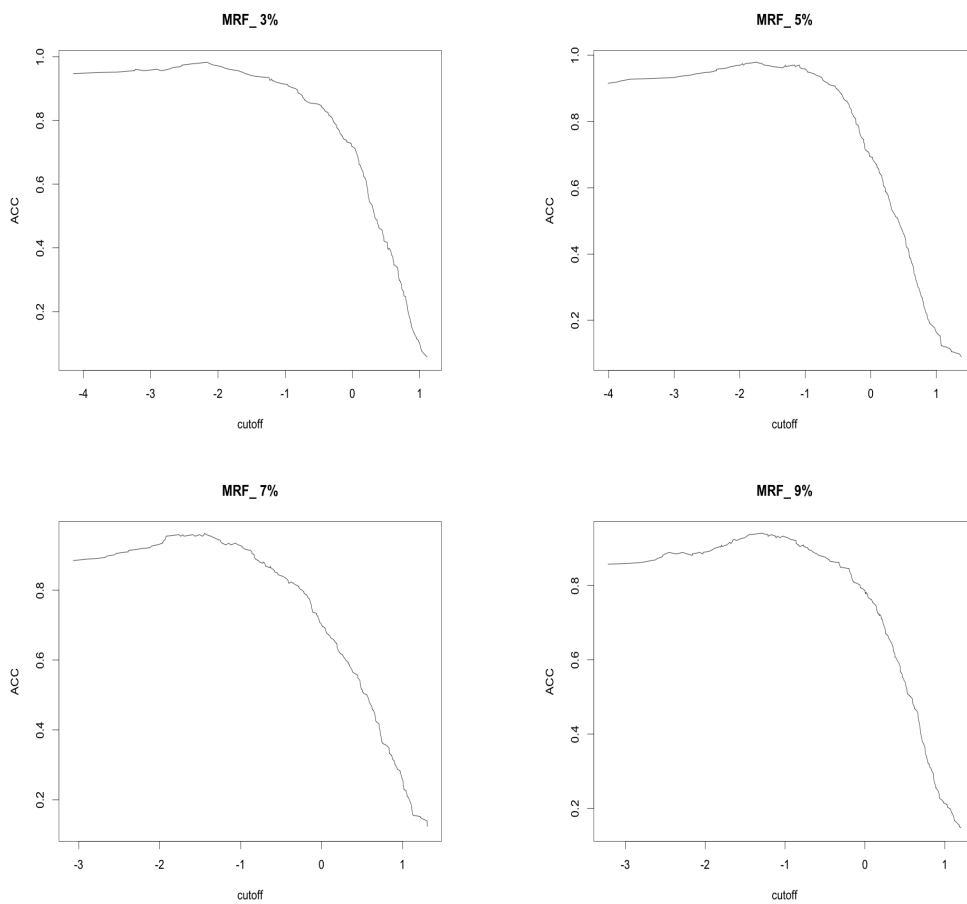


그림 14: 총 30문항에 대한 컷오프 기준에 따른 ACC

세부 연구3. 부분 비정상 반응 패턴에 대한 연구

모의실험 1. 무선 문항 및 무선 반응 조작

그림 15와 그림 16에서 최적의 컷오프 기준은 조건에 따라 미세하게 변화하지만, 여러 조건들을 동시에 고려하였을 때 대체로 높은 정확도를 보이는 $-3 \sim -2$ 가 최적의 컷오프 값을 알 수 있다.

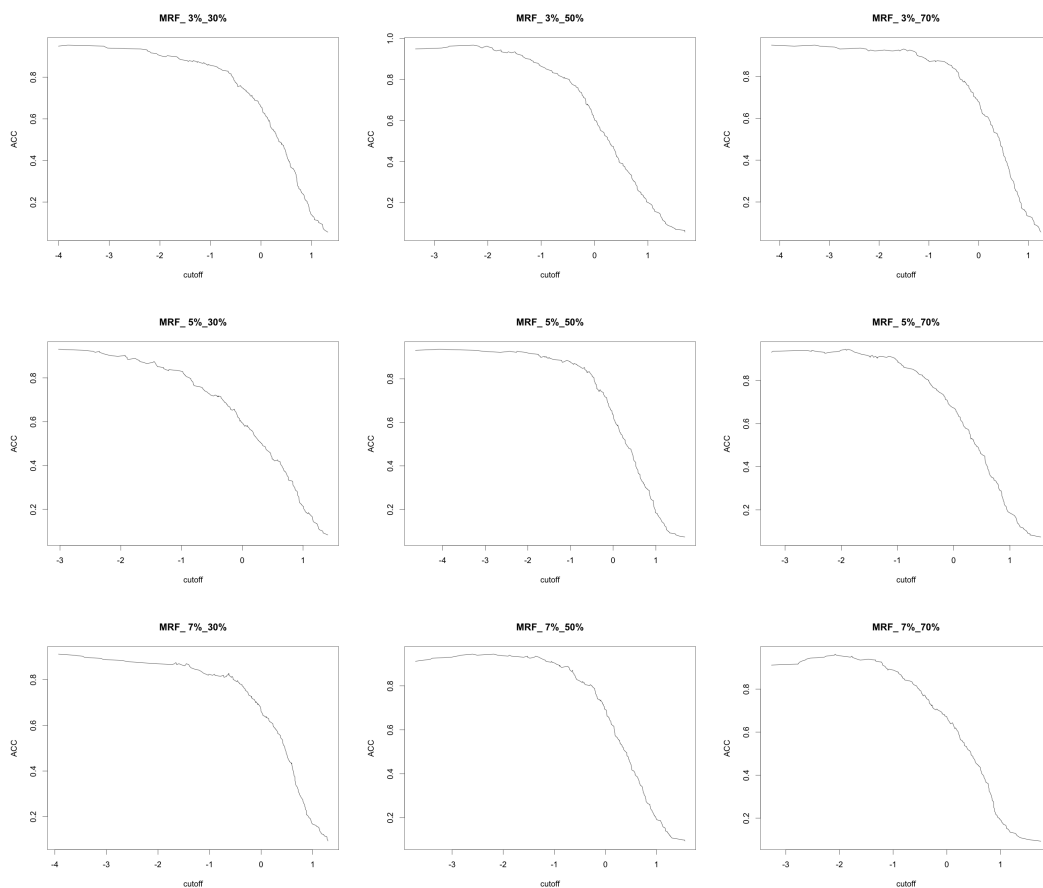


그림 15. 모의실험1(무선 문항 및 무선 반응 조작)의 ACC.

MRF_3%_30% 에서 앞의 수치(3%)는 비정상비율, 뒤의 수치(30%)는 조작한 문항 비율을 나타낸다.

모의실험 2. 연속 문항 및 연속 반응 조작

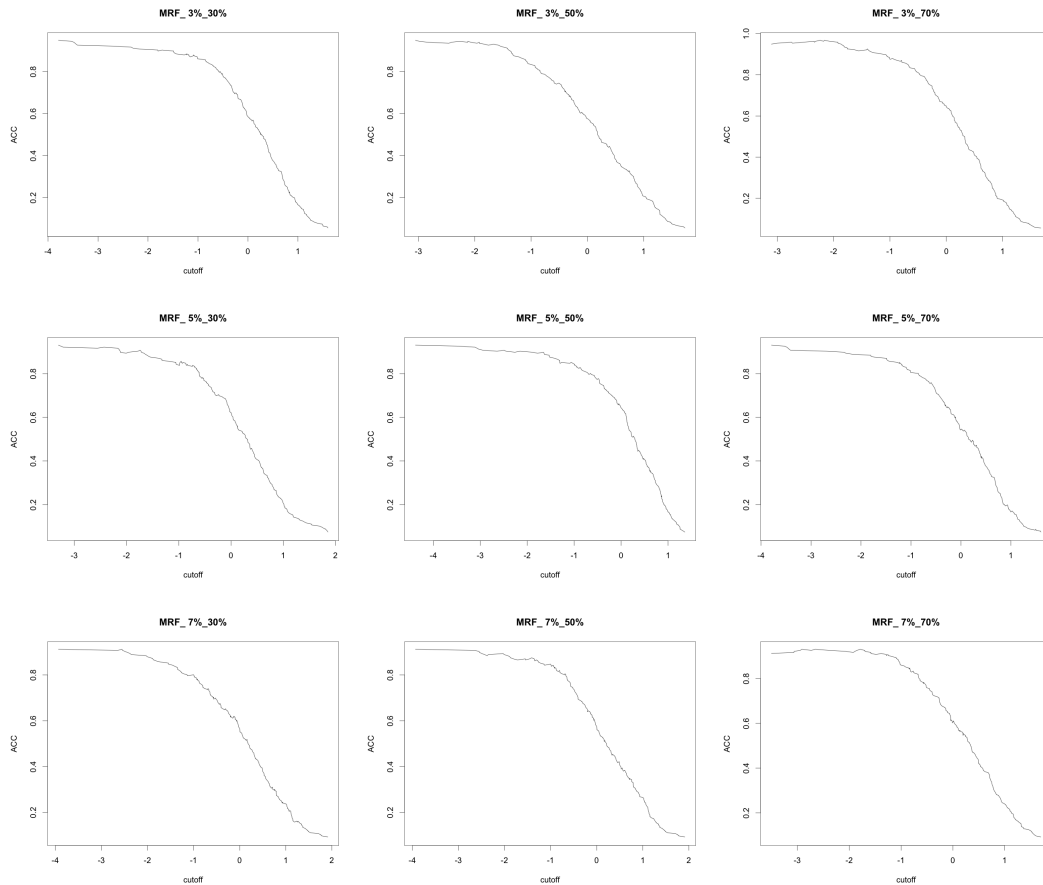


그림 16. 모의실험 2(연속 문항 및 연속 반응 조작)의 ACC.

MRF_3%_30% 에서 앞의 수치(3%)는 비정상비율, 뒤의 수치(30%)는 조작한 문항 비율을 나타낸다.

컷오프 기준에 대한 연구를 종합하면 부분 문항에 대한 비정상 반응 패턴의 경우 최적의 컷오프 값은 -2 ~ -3으로 나타났지만, 전체 문항에 대한 비정상 반응 패턴의 경우 -3 부근에서는 오히려 정확도가 감소하므로 최종적으로 -2를 컷오프 값으로 결정하였다.

5.2.3. 컷오프 기준을 이용한 모형 비교

정상 반응 패턴 225개와 비정상 반응 패턴 225개 중 VRIN 문항과 TRIN 문항에 결측치(missing value)가 있는 것을 제외한 정상 반응 패턴 208개와 비정상 반응 패턴 187개로 컷오프 기준에 따라 VRIN, TRIN, IRT, MRFs 모형의 정확도를 구하였다.

한국판 MMPI-2 매뉴얼의 기준에 따라 VRIN과 TRIN의 T점수가 80점 이상인 반응 패턴을 비정상 반응 패턴으로 판별하였다. MRFs와 IRT는 1) 표준화된 로그 우도를 -2를 기준으로 하여 판별하고, 2) 로그 우도를 $\log(.5) \times I$ (총 문항수)를 기준으로 판별하였다. 앞서 AUC와 달리 training set과 test set을 나누지 않고 분석하였다.

표 12와 표 13을 보면 왜곡지표 VRIN과 TRIN의 경우 정확도(ACC)가 낮으며, 표 14와 표 15를 보면 IRT의 l_z 와 l_0 역시 정확도(ACC)가 낮다. IRT의 l_z 는 -2를 기준으로 l_0 는 $\log(.5) \times I$ 를 기준으로 판별하였다.

표 12. VRIN의 Confusion Matrix

VRIN		실제 라벨	
		비정상	정상
예측	비정상	32	0
라벨	정상	155	208
ACC		.607	
(95% 신뢰구간)		(.557, .656)	

표 13. TRIN의 Confusion Matrix

TRIN		실제 라벨	
		비정상	정상
예측	비정상	9	0
라벨	정상	178	208
ACC		.549	
(95% 신뢰구간)		(.499, .599)	

표 14. IRT(l_z)의 Confusion Matrix

IRT(l_z)		실제 라벨	
		비정상	정상
예측	비정상	51	0
라벨	정상	136	208
ACC		.656	
(95% 신뢰구간)		(.606, .703)	

표 15. IRT(l_0)의 Confusion Matrix

IRT(l_0)		실제 라벨	
		비정상	정상
예측	비정상	61	0
라벨	정상	126	208
ACC		.681	
(95% 신뢰구간)		(.633, .727)	

표 16. MRFs의 Confusion Matrix

MRFs		실제 라벨	
		비정상	정상
예측 라벨	비정상	168	0
	정상	19	208
ACC		.952	
(95% 신뢰구간)		(.925, .971)	

반면, 표 16의 MRFs 모형의 경우 $\log(.5) \times I$ 를 컷오프 기준으로 사용하면, VRIN, TRIN, IRT 보다 상대적으로 높은 정확도를 보인다. 그림 17의 히스토그램에서 점선은 무선 반응 패턴들의 로그 우도이고, 실선은 정상 반응 패턴들의 로그 우도이다. 가운데 기준선은 $\log(.5) \times 567$ 이다. 그러나, 그림 17 에서 알 수 있듯이 비정상 반응 패턴의 비율이 많으면 -2의 컷오프 기준을 사용하기 어렵다. 왜냐하면, 비정상 반응 패턴은 확률적으로 발생하기 어려운 반응 패턴임에도 불구하고, 다른 이유 등에 의하여 비정상 반응 패턴이 많을 경우 -2의 컷오프 기준보다 더 높은 값에서 형성되기 때문이다.

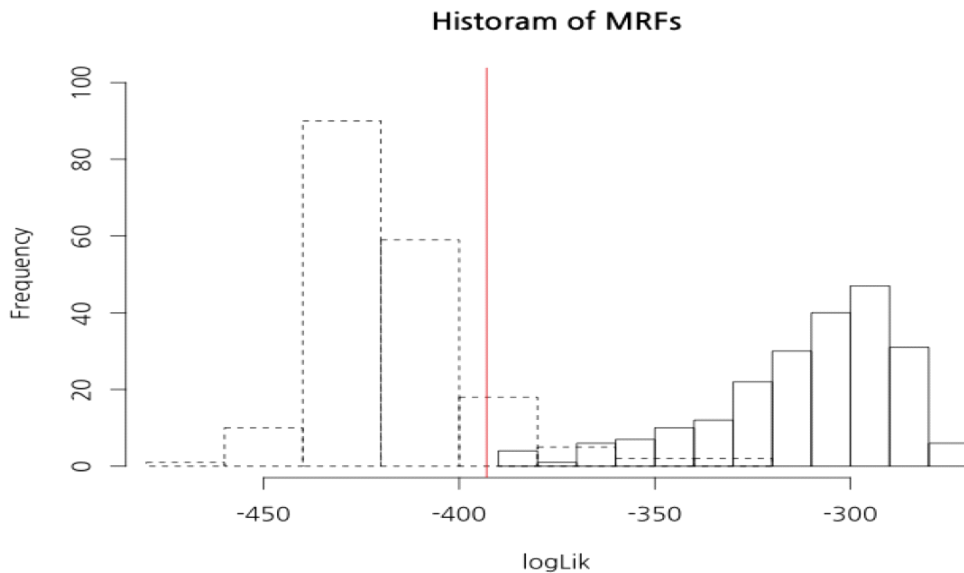


그림 17: 정상 반응 패턴과 비정상 반응 패턴들에 대한 MRFs 모형에서의 로그 우도 히스토그램

세부 연구1. 비정상 패턴 비율 조정

다음으로 앞의 AUC 분석과 마찬가지로 비정상 반응 비율을 25%, 20%, 15%, 10%에 대하여 VRIN, TRIN, IRT, MRFs의 ACC를 비교하였다.

표 17: VRIN, TRIN, IRT, MRFs의 평균 ACC(총 567문항)
(괄호 안은 표준편차, 총 10회 시행)

비정상 반응 비율	Base ACC	VRIN	TRIN	왜곡 지표 유무 포함	IRT	MRFs
25%	.751	.782 (.011)	.760 (.009)	(567문항)	.965 (.007)	.983 (.003)
				제거 (456문항)	.929 (.009)	.987 (.002)
20%	.8	.878 (.007)	.866 (.012)	(567문항)	.966 (.014)	.976 (.004)
				제거 (456문항)	.970 (.004)	.988 (.003)
15%	.849	.872 (.009)	.857 (.012)	(567문항)	.976 (.014)	.986 (.001)
				제거 (456문항)	.982 (.011)	.985 (.006)
10%	.9	.924 (.008)	.905 (.004)	(567문항)	.967 (.015)	.976 (.003)
				제거 (456문항)	.969 (.009)	.978 (.002)

표 17에서 Base ACC는 모든 비정상 반응 패턴을 정상 반응 패턴으로 분류하였을 경우의 ACC이다. 분석 결과, 모든 조건에서 MRFs의 ACC가 IRT의 ACC보다 높고, IRT와 MRFs는 왜곡지표(VRIN, TRIN)보다 ACC가 더 높음을 알 수 있다.

표 18을 보면 B. ACC 역시 MRFs 는 VRIN, TRIN, IRT 보다 높은 정확도를 보였고, IRT 역시 VRIN과 TRIN 보다 더 높은 정확도를 보였다. B. ACC로 분석한 결과는 ACC로 분석한 결과와 동일하다.

표 18: VRIN, TRIN, IRT, MRFs의 평균 B. ACC(총 567문항)
(괄호 안은 표준편차, 총 10회 시행)

비정상 반응 비율	Base B. ACC	VRIN	TRIN	왜곡 지표 유무 포함	IRT	MRFs
25%	.375	.583 (.027)	.527 (.028)	(567문항)	.931 (.014)	.981 (.007)
				제거 (456문항)	.858 (.019)	.984 (.004)
20%	.4	.579 (.032)	.526 (.023)	(567문항)	.866 (.012)	.973 (.005)
				제거 (456문항)	.931 (.014)	.988 (.005)
15%	.425	.583 (.02)	.540 (.02)	(567문항)	.973 (.012)	.983 (.005)
				제거 (456문항)	.974 (.015)	.987 (.007)
10%	.45	.632 (.044)	.542 (.021)	(567문항)	.967 (.020)	.976 (.018)
				제거 (456문항)	.964 (.022)	.984 (.009)

표 19과 표 20은 비정상 반응 패턴 비율이 각각 9%, 7%, 5%, 3%일 때 ACC를 나타낸 것이다. 표 19과 표 20에서 MRFs(1)은 표준화된 개인의 로그 우도(l_z)를 -2 기준으로 판별한 것이고, MRFs(2)는 개인의 로그 우도(l_0)를 $\log(.5) \times \text{문항수}$ 로 판별한 것이다.

표 19: 비정상 반응 패턴 비율에 따른 VRIN, TRIN, IRT, MRFs의 평균 ACC
(총 567문항)

비정상 반응 패턴 비율	Base ACC	VRIN	TRIN	IRT	MRFs (1)	MRFs (2)
9%	.908	.919 (.004)	.913 (.001)	.973 (.021)	.996 (.005)	.975 (.005)
7%	.928	.939 (.003)	.935 (.003)	.974 (.012)	.997 (.003)	.976 (.003)
5%	.949	.968 (.001)	.951 (.011)	.927 (.003)	.991 (.001)	.978 (.001)
3%	.972	.976 (.012)	.973 (.011)	.91 (.006)	.978 (.001)	.978 (.001)

표 20: 비정상 반응 패턴 비율에 따른 VRIN, TRIN, IRT, MRFs의 평균 B. ACC
(총 567문항)

비정상 반응 패턴 비율	Base B.ACC	VRIN	TRIN	IRT	MRFs (1)	MRFs (2)
9%	.455	.560 (.011)	.524 (.012)	.976 (.004)	.976 (.034)	.964 (.034)
7%	.465	.563 (.002)	.563 (.002)	.972 (.014)	.984 (.022)	.973 (.022)
5%	.475	.682 (.002)	.485 (.011)	.961 (.001)	.995 (.001)	.988 (.001)
3%	.485	.572 (.001)	.495 (.013)	.954 (.003)	.988 (.001)	.988 (.001)

세부 연구2. 적은 문항(30문항)

왜곡지표를 제외한 30문항을 샘플링하여 VRIN과 TRIN, IRT, 그리고 MRFs의 ACC 및 B. ACC를 비교하였다. 총 10회 반복 수행하였다.

표 21: 비정상 반응 패턴 비율에 따른 IRT와 MRFs의 평균 ACC 및 B. ACC
(총 30문항), (100회 반복시행, 괄호안은 표준편차)

비정상 반응 패턴 비율	ACC			B. ACC		
	IRT	MRFs (1)	MRFs (2)	IRT	MRFs (1)	MRFs (2)
9%	.950 (.012)	.970 (.012)	.956 (.049)	.729 (.082)	.855 (.049)	.954 (.064)
7%	.961 (.001)	.977 (.024)	.974 (.011)	.776 (.002)	.920 (.077)	.986 (.006)
5%	.972 (.007)	.965 (.036)	.972 (.002)	.798 (.098)	.887 (.028)	.970 (.018)
3%	.973 (.009)	.956 (.015)	.982 (.012)	.922 (.074)	.917 (.124)	.936 (.099)

표 21 에서 볼 수 있듯이 왜곡지표를 제외한 30문항 역시 MRFs가 IRT에 비해 상대적으로 높은 정확도를 보인다. 그러나, 비정상 비율이 5%와 3%의 경우에 IRT의 경우가 MRFs 경우보다 ACC는 더 높고, B. ACC는 더 낮다. 그 이유는 IRT가 비정상 반응 패턴을 분류를 잘 못한 대신 상대적으로 정상 반응 패턴을 잘 분류하였기 때문이다.

논의

연구 1을 종합해보면, 왜곡지표(VRIN과 TRIN)를 포함한 567문항을 모두 사용한 경우와 왜곡지표를 제외한 456문항을 사용한 경우 비정상 반응 비율이 5% 이상에서는 MRFs가 IRT에 비해 AUC, ACC, B. ACC 모든 면에서 더 좋은 성능을 보였다. 왜곡지표를 제외한 30문항에서도 5% 이상의 경우 MRFs가 IRT에 비해 더 좋은 성능을 보였다.

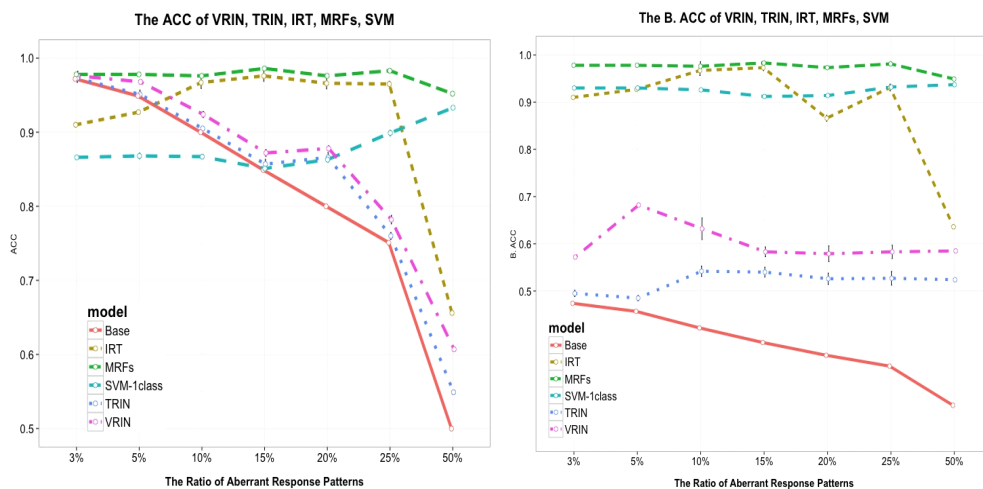


그림 18: 비정상 반응 패턴 비율에 따른 IRT, MRFs, SVM, TRIN, VRIN의 ACC(왼쪽) 및 B. ACC(오른쪽). (총 567문항)

그림 18을 보면, SVM one(1) class³⁾ 모형을 기준으로 하였을 때, 비정상 패턴이 20% 이상이 되면 IRT의 정확도는 낮아진다. 그러나, MRFs의 경우 비정상 반응 패턴 비율과 관계없이 IRT를 비롯한 다른 모형들보다 높은 정확도를 보이며, 표 22를 보면 문항이 적은 30문항에서도 MRFs는 높은 성능을 보인다.

또한, ACC와 B. ACC 역시 AUC 분석과 마찬가지로 567문항 혹은 30문항을 training set과 test set을 나누어 분석하였을 경우에도 MRFs가 IRT 보다 더 높은 성능을 보였다(김형준 등, 2014).

3) 정상 반응 패턴만을 가지고 학습하였으며, nu 파라미터는 0 ~ 0.1 까지 100 구간을 나누어 training 오류(error)를 최소화하는 파라미터를 선택 하였다.

6. 연구 2

연구 2에서는 마코프랜덤필드(MRFs) 모형의 응용 가능성을 검증한다. 설문지의 길이가 길어지면 연구 참여자는 피로가 증가하고 흥미가 감소할 수 있다. 따라서, 설문지의 길이가 길어지면 다양한 측면에서 응답의 질이 감소하게 된다. 설문지 길이가 길어질수록 한 줄로 반응하는 응답하는 비율이 증가(Deutskens et al., 2014)하고, ‘모르겠다’와 같은 응답의 비율이 증가한다(Herzog & Bacman, 1981). 또한, 주관식 질문에 대한 응답의 길이가 짧아지고, 무응답의 비율이 증가한다(Galesic & Bosnjak, 2009).

김권현 등(2014)은 문항의 위치에 따라, 즉 설문이 길어질수록 1) 개인별 엔트로피가 감소하고, 개인별 최대 한줄 응답 길이가 증가하고, 램펠-지브 복잡도가 감소함을 보였다. 김권현 등(2014)의 연구는 설문지 길이에 따라 사람들이 전반적으로 응답의 질이 나빠짐을 밝히기 위한 연구이다. 따라서, 본 논문에서 제안하는 MRFs 모형으로는 전반부와 후반부 혹은 초반, 중반, 중반 3등분으로 구분하여 전체 사람들의 우도의 평균을 비교하였지만, 설문 길이에 따른 전체 사람들의 우도는 일관되게 차이가 나지 않았다.

그러나, 본 논문에서는 전체 응답의 질이 감소하는 것을 밝히는 것이 목적이 아니라 비정상 반응 패턴의 비율이 증가하는 것이 연구주제이므로, 1) 표준화된 개인의 로그 우도(l_z)를 -2 기준으로, 2) 개인의 로그 우도(l_0)를 $\log(.5) \times \text{문항수}$ 기준으로 하여 비정상 반응 패턴들의 비율이 설문지가 길어짐에 따라 증가하는지를 연구해보았다.

6.1. 설문자료 및 자료분석방법

6.1.1. 설문자료

한국 리서치에서 2013년 웹을 통해 실시한 두 개의 설문자료를 분석하였다. 첫 번째 자료는 기업 이미지에 관한 조사로서, 설문소요시간의 중앙값이 21분이었고, 참가 인원은 1300명이었다. 두 번째 자료는 상품에 대한 이용 실태에 관한 조사로서, 설문소요시간의 중앙값은 51분이었고, 참가인원은 5000명이었다.

이 자료 중에서 리커트 응답만을 선별하여 분석하였다. 첫 번째 자료는 5점 척도(scale)로서 총 268문항이었고, 두 번째 자료는 7점 척도로 총 249문항이었다. 두 자료 모두 결측치는 없었다. 그러나, 두 번째 자료는 분기 질문이 존재하여 응답자별로 결측치가 존재한다. 분기 질문이란 특정한 응답을 하는 사람에게만 그 다음 질문이 제시되는 문항을 말한다. 예를 들어, ‘최근 일주일 내에 영화를 보셨습니까?’ 라는 문항에서 ‘예’ 라고 응답한 사람에게만 ‘영화관에서 보셨습니까?’ 라는 다음 질문이 제시되는 문항을 말한다. 설문자료 중 분기 질문을 제외하고 총 113문항을 분석에 사용하였다. 편의상 첫 번째 자료를 ‘짧은 설문지’, 두 번째 자료를 ‘긴 설문지’로 부르겠다.

본 설문자료는 연구자가 측정하고자 하는 특성이 여러 차원으로 서로 혼재되어 있으며, 어떤 문항들이 어떤 특성과 연결되어 있는지 파악하기 힘들다. 따라서, 문항반응이론으로 비정상 반응패턴을 탐지하기는 힘들다. 그러나, MRFs 모형의 경우 특별한 가정을 요구하지 않고 주어진 반응 패턴들만으로 모형화하기 때문에, 문항의 특성들이 혼재된 경우에도 비정상 반응 패턴을 탐지하는 데 이용할 수 있다.

6.1.2. 자료분석방법

자료의 분석을 위해 크게 2 가지 방법을 사용하였다. 첫째, 전체 문항 중에서 2등분(전반부, 후반부) 혹은 3등분(초반, 중반, 종반) 하여 각 부분에서 비정상 반응 패턴이 증가하는 정도를 알아보았다. 둘째, 이동창(moving window)으로 문항을 선택하여 문항의 위치에 따른 비정상 반응 패턴의 증가를 살펴보았다.

본 연구에서 제안하는 MRFs 모형은 이진(binary) 모형이기 때문에 설문지 1은 5점 척도로서 3점 이상의 값을 1로 3점 미만의 값을 0으로, 설문지 2는 7점 척도로서 5점 이상의 값을 1로 5점 미만의 값을 0으로 재코딩하여 분석하였다.

MRFs를 이용하여 전체 반응 패턴에 대하여 포텐셜 파라미터를 추정 한 후, 각 부분(등분 혹은 이동창)별로 표준화된 개인의 로그 우도를 이용하여 분석하였다.

6.1.3. 분석결과 및 논의

분석결과

우선, ‘짧은’ 설문지의 경우 표 22와 표 23을 보면 전체 문항을 각각 이등분 혹은 삼등분 하였을 때 비정상 반응 패턴의 증가는 비일관된 결과를 보인다.

표 22: 짧은 설문지에서의 전반부, 후반부의 비정상 반응 패턴의 수
(괄호안은 비율, / 는 문항수)

	전반부	후반부	
$l_0 \leq \log(.5) \times I$	10(.008)	5(.008)	$\chi^2(1) = 1.073$ $p = .300$
$l_z \leq -2$	63 (.048)	57(.043)	$\chi^2(1) = .218$ $p = .640$

표 23: 짧은 설문지에서의 초반부, 중반부, 종반부의 비정상 반응 패턴의 수.
(괄호안은 비율, / 는 문항수)

	초반부	중반부	종반부	
$l_0 \leq \log(.5) \times I$	17(.013)	18(.013)	6(.005)	$\chi^2(2) = 6.56$ $p = .04$
$l_z \leq -2$	58 (.045)	62 (.047)	74 (.057)	$\chi^2(2) = 2.26$ $p = .323$

둘째, 이동창(moving window)으로 문항을 선택하여 문항의 위치에 따른 비정상 반응 패턴의 증가를 살펴보았다. 윈도우의 크기는 50으로 고정하였다. 아래 그림 19 에서도 짧은 설문지에서는 뒤로 갈수록 비정상 반응 패턴의 수가 증가하지 않는다.

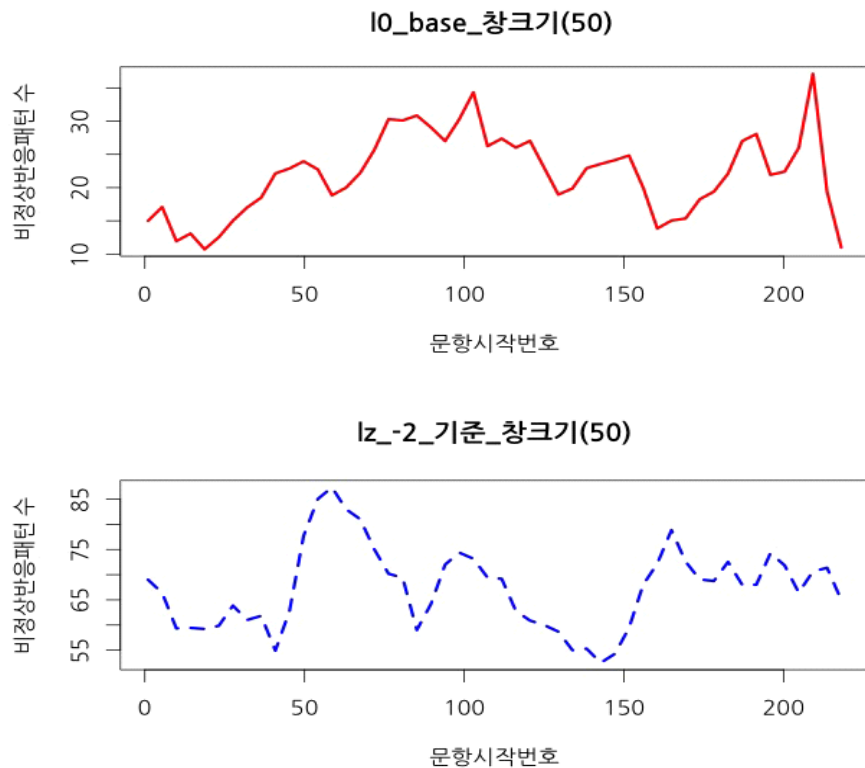


그림 19: 짧은 설문지에서 문항 위치에 따른 비정상 반응 패턴 수 (윈도우사이즈는 50). x 축의 문항시작번호는 이동창(moving window) 시작 위치를 말한다. 위의 그래프는 $l_0 \leq \log(.5) \times \text{문항수}$ 기준으로 판별한 그래프이고, 아래 그래프는 $l_z \leq -2$ 로 판별한 그래프이다.

반면, 긴 설문지에서는 비정상 반응 패턴의 비율이 시간이 지날수록 더 크게 증가함을 알 수 있다. 이등분을 하였을 때, 표 24에서 보면 전반부 보다 후반부의 비정상 반응 패턴이 증가하였으며, 표 25에서처럼 삼등분을 하였을 때 역시 뒤로 갈수록 비정상 반응 패턴의 증가를 보였다.

표 24: 긴 설문지에서의 전반부, 후반부의 비정상 반응 패턴의 수.
(괄호안은 비율, / 는 문항수)

	전반부	후반부	
$l_0 \leq \log(.5) \times I$	376(.075)	504(.101)	$\chi^2(1) = 20.097$ $p < .001$
$l_z \leq -2$	132 (.027)	391(.08)	$\chi^2(1) = 134.3$ $p < .001$

표 25: 긴 설문지에서의 초반부, 중반부, 종반부의 비정상 반응 패턴의 수.
(괄호안은 비율, / 는 문항수)

	초반부	중반부	종반부	
$l_0 \leq \log(.5) \times I$	441(.088)	447(.089)	529(.11)	$\chi^2(2) = 11.30$ $p < .001$
$l_z \leq -2$	191(.038)	250(.05)	348(.07)	$\chi^2(2) = 46.83$ $p < .001$

또한, 그림 20을 보면 긴 설문지에서는 이동창(moving window)의 위치에 따라 비정상 반응 패턴이 증가하는 경향을 보였다.

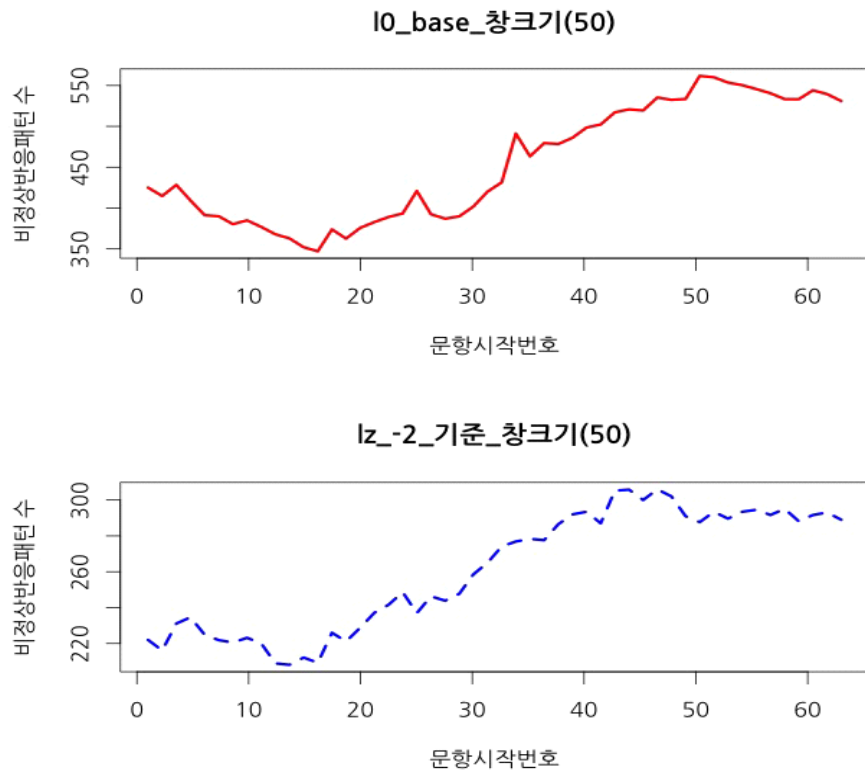


그림 20: 긴 설문지에서 문항 위치에 따른 비정상 반응 패턴 수 (원도우사이즈는 50). x 축의 문항시작번호는 이동창(moving window) 시작 위치를 말한다. 위의 그래프는 $l_0 \leq \log(.5) \times \text{문항수}$ 기준으로 판별한 그래프이고, 아래 그래프는 $lz \leq -2$ 로 판별한 그래프이다.

논의

위 결과는 김권현 등(2014)의 연구 결과와 일치한다. 짧은 설문지에서는 비정상 반응 패턴들이 증가하지 않았지만, 긴 설문지에서는 비정상 반응 패턴들이 증가하였다. 김권현 등(2014)의 연구에서도 역시 짧은 설문지에서는 응답의 질이 다소 일정하게 유지된 반면, 긴 설문지에서는 응답의 질이 저하되었다. 두 연구를 종합해보면, 설문지의 길이가 길어짐에 따라 전체 사람들의 응답 질이 저하되는 것과 더불어 비정상 반응 패턴의 비율 역시 증가함을 알 수 있다.

짧은 설문지의 경우 비정상 반응 패턴들이 증가하지 않았는데, 약 20분정도 소요되는 짧은 설문지의 경우 설문의 길이 효과가 발생하기에는 짧은 시간이었음을 보여준다.

7. 결론

본 연구에서는 비정상 반응 패턴을 탐지하기 위해 MRFs 모형을 제안하였다. MRFs 모형에서 포텐셜 함수는 MMPI-2의 왜곡지표인 VRIN과 TRIN의 속성을 반영하도록 구성하였다. 연구 1의 결과 왜곡지표(VRIN, TRIN)와 더불어 기존의 문항반응이론 보다 여러 조건들에서 더 높은 정확도를 보였다. 별도의 문항을 사용하지 않고도 모형만으로 비정상 반응 패턴을 탐지할 수 있다는 것이 본 연구의 가장 큰 시사점이다.

연구 2에서는 설문지의 길이에 따라 전반적인 응답의 질이 떨어짐을 보인 선행연구 결과에 덧붙여, MRFs 모형을 통해 설문지의 길이가 길어짐에 따라 비정상 반응 패턴이 증가할 수 있음을 보였다.

본 연구의 의의를 정리하면 다음과 같다. 1) 별도의 모형 없이, 그리고 모형에 대한 별다른 가정 없이도 비정상 반응 패턴을 탐지할 수 있는 MRFs 모형을 제안하였고 그 변별성을 검증하였다. 2) 본 연구에서 제안하는 MRFs 모형의 비정상 반응을 잘 판단할 수 있는 기준을 제시하였다. 3) 비정상 반응 패턴을 판별하여 설문지 길이에 따라 비정상 반응 패턴이 증가할 수 있음을 보였다. 또한, 4) 본 모형은 VRIN과 TRIN의 속성을 가지고 있기 때문에, VRIN과 TRIN의 문항이 포함되면 비정상 반응 패턴을 더 잘 분류할 수 있다.

그러나, 본 연구에서의 한계점 역시 존재한다. 첫째, 본 연구에서는 비정상 반응 패턴 탐지를 위해 오직 문항 반응만을 사용하고 있다. 설문지에서 단순히 문항 반응뿐만 아니라 반응 속도 및 마킹의 강도와 크기 등을 특질(feature)로 활용한다면 더 높은 정확도를 보일 수 있다. 실제로, 연구 1의 경우에서도 IRT로 구한 참여자의 특성(θ)을 0을 기준으로 하여, 이진화 한 뒤 조건부랜덤필드 모형을 사용한 경우 특질(feature)이 없는 마코프랜덤필드에 비해 좀 더 높은 정확도를 보였다. 또한, 참가자의 기분, 성별 등의 변수들을 고려하였을 때 반응 패턴의 확률들이 어떻

게 달라지는 지 등의 응용 연구도 필요하다.

둘째, 본 연구 모형에서는 이진 변수만을 사용하고 있는데, 보통의 설문지는 연속적인 변수도 존재하기 때문에 연속 함수를 다룰 수 있는 모형으로 확장할 필요성이 있다. 즉, 상태에 대하여 단절적(discrete)인 오류 함수를 정의하기보다, 거리에 따라 오류 함수로 재정의 할 필요성이 제기된다.

끝으로, 본 연구 모형에서는 은닉 변수(hidden variable)가 존재하지 않는 모형을 사용하였다. 본 모형에서 가우시안 은닉(hidden) 변수를 사용하고 관찰된(observed) 문항 반응과 은닉 변수 사이의 함수를 로지스틱 함수를 사용한다면, 기존 IRT의 2PL 모형에 전이 함수를 더한 모형으로 확장 할 수 있을 것이다. 이 때, 노드 포텐셜은 난이도와 상응하고, 은닉 변수와 관찰된 문항 반응과의 엣지 포텐셜은 변별도와 상응할 수 있다.

참 고 문 헌

- 김권현·유동주·김형준·김청택. 2014. “설문지의 길이가 응답의 질에 미치는 영향.” 심사중.
- 김형준·김청택·장병탁. 2014. “마코프랜덤필드를 이용한 설문지에서의 비정상 반응 탐지.” 《한국정보과학회 동계학술발표회 논문집》. 41: 577-579.
- 장은경. 2010. 《반응자 적합도와 마코프 체인을 이용한 한국판 MMPI-2의 반응왜곡 탐지》. 서울대학교 석사학위 논문.
- Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer.
- Bernal, A., Crammer, K., Hatzigeorgiou, A., & Pereira, F. (2007). “Global discriminative learning for higher-accuracy computational gene prediction.” *PLoS Computational Biology*. 3(3).
- Bhatt, M. R. and Desai, U. B. (1994), “Robust image restoration algorithm using Markov random field model.” *CVGIP : Graphical Models and Image Processing*, vol. 56: 61-74.
- Butcher, J., Dahlstrom, W., Graham, J., Tellegen, A., & Kaemmer, B. (1989). *Minnesota multiphasic personality inventory-2 (MMPI-2) : Manual for administration and scoring*. Minneapolis: University of Minnesota Press.
- Butcher, J., & Han, K. (1995). “Development of an MMPI-2 scale to assess the presentation of self in a superlative manner: The S scale.” *Advances in personality assessment*, 10: 25-50.
- Chandgotia, N., Han, G., Marcus, B., Meyerovitch, T., and Pavlov, R. (2014). “One dimensional markov random fields, markov chains and topological markov random fields.” *Proceedings of American Mathematical Society*. 142: 227-242.
- Curran, P. G., Kotrba, L., & Denison, D. (2010). “Careless responding in surveys: Applying traditional techniques to organizational settings.”

Paper Presented at the 25th Annual Conference of the Society for Industrial/Organizational Psychology.

- Daskalaki, S., Kopanas, I., Avouris, N. (2006). "Evaluation of classifiers for an uneven class distribution problem." *Applied Artificial Intelligence* 20. 381-417.
- Deutskens, E., De Ruyter, K., Wetzels, M., and Oosterveld, P. 2004. "Response rate and response quality of internet-based surveys: an experimental study." *Marketing letters* 15(1): 21-36.
- Drasgow, F., Levine, M. V., and Williams, E. A. (1985). "Appropriateness measurement with polychotomous item response models and standardized indices." *British Journal of Mathematical and Statistical Psychology* 38(1): 67-86.
- Ehlers, C., Greene-Shortridge, T. M., Weekley, J. A., & Zajack, M. D. (2009). "The exploration of statistical methods in detecting random responding." *Paper presented at the annual meeting of the Society for Industrial/Organizational Psychology.*
- Elazmeh, W., Japkowicz, N., Matwin, S. (2006). "Evaluating misclassifications in imbalanced data." *Proceedings 17th European Conference on Machine Learning: ECML 2006*, 126 - 137.
- Fawcett, T. (2006). "An introduction to ROC analysis." *Pattern recognition letters*. 27(8): 861-874.
- Ferrando, P., & Chico, E. (2001). "Detecting dissimulation in personality test scores: A comparison between person-fit indices and detection scales." *Educational and Psychological Measurement*. 61(6): 997-1012.
- Galesic, M., & Bosnjak, M. (2009). "Effects of questionnaire length on participation and indicators of response quality in a web survey." *Public Opinion Quarterly* 73(2): 349-360.
- Gao, F., & Chen, L. (2005). "Bayesian or non-Bayesian: A comparison study of item parameter estimation in the three-parameter logistic model." *Applied Measurement in Education*, 18 (4), 351-380.
- Hand, D. J., & Till, R. J. (2001). "A simple generalisation of the area

- under the ROC curve for multiple class classification problems.” *Machine Learning*, 45(2): 171 - 186
- Hanson, B. A., & Beguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, 26 (1), 3-24.
- Herzog, A., & Bachman, J. (1981). “Effects of questionnaire length on response quality.” *Public Opinion Quarterly* 45(4): 49-59.
- Lafferty, J., McCallum, A., & Pereira, F. (2001). “Conditional random fields: Probabilistic models for segmenting and labeling sequence data.” *International Conference on Machine Learning (ICML)*
- Lange, K. (2010). *Numerical Analysis for Statisticians*. New York : Springer
- Levine, M. V., and Rubin, D. B. (1979). “Measuring the appropriateness of multiple choice test scores.” *Journal of Educational Statistics* 4(4): 269-290.
- Ling, C., Huang, J., and Zhang, H. (2003). “AUC: A statistically consistent and more discriminating measure than accuracy.” *Proceedings of International Joint Conference on Artificial Intelligence*.
- Liu, D. and Nocedal, J. (1989). “On the limited memory method for large scale optimization.” *Math. Program. B* 45(3): 503-528.
- Murphy, K. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.
- Nichols, D., Greene, R., & Schmolck, P. (1989). “Criteria for assessing inconsistent patterns of item endorsement on the MMPI: Rationale, development, and empirical trials.” *Journal of clinical psychology*. 45(2), 239-250.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.
- Reise, S. P. and Due, A. (1991). “The influence of test characteristics on the detection of aberrant response patterns.” *Applied Psychological Measurement* 15(3): 217-226.

- Rother, C., Kolmogorov, V. & Blake, A. (2004) "Grabcut: Interactive foreground extraction using iterated graph cuts." *ACM Transactions on Graphics (SIGGRAPH)*. 23(3): 309-314.
- Quattoni, A., Wang, S., Morency, Collins, M., & Darrell, T. (2007). "Hidden state conditional random fields." *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Wu, L. Y (2014). *CRF: R code for undirected graphical models*
- Spackman, K. A., (1989). "Signal detection theory: Valuable tools for evaluation inductive learning." *Proceedings of Sixth Internat. Workshop on Machine Learning*. Morgan Kaufman, San Mateo, CA, 160-163.
- Schmitt, N., Chan, D., Sacco, J. M., McFarland, L. A., & Jennigs, D. (1999). "Correlates of Person Fit and Effect of Person Fit on Test Validity." *Applied Psychological Measurement*. 23(1): 41-53
- Schmidt, M. (2012). *Ugm: Matlab code for undirected graphical models*
- Weisberg, H. F. (2009). *The total survey error approach: A guide to the new science of survey research*. University of Chicago Press.
- Zhao, Y., Zhuang, X., L. and Anderson, L. (1992), "Parameter estimation and restoration of noisy image using Gibbs distributions in hidden Markov models." *CVGIP: Graphical Models and Image Processing*, 54: 187-197.
- Zhu, J. & Xing, E. (2010). "Conditional topic random fields." *The 27th International Conference on Machine Learning (ICML)*.

Abstract

Detection of Aberrant Response on a Questionnaire Using MRFs

Hyung-Jun Kim

Interdisciplinary Program in Cognitive Science

The Graduate School

Seoul National University

The main purpose of this study is the detection of aberrant response patterns such as random response, using Markov Random Fields(MRFs). MRFs are able to include the information of item characteristics on the node potential and transitions of responses on the edge potential.

To determine the performance of MRFs, the AUC(Area Under the Curve) of Receiver Operating Characteristic(ROC) curve was used in study 1. MRFs can consistently detect the aberrant response patterns with all conditions in MMPI-2 questionnaire than Variable Response Inconsistency(VRIN), True Response

Inconsistency(TRIN), Item Response Theory(IRT), and Support Vector Machine(SVM). However, the exact cut-off value of MRFs is required to classify which of the aberrant response patterns in whole response patterns. Cut-off value is considered -2 or $\log(.5) \times I$ ($I = \text{total number of items}$) that is the base log-likelihood of a whole random response pattern. MRFs also outperform the prediction of aberrant response patterns in accuracy.

In study 2, MRFs is used to find out the increase rate of aberrant responses with the questionnaire length. To examine the hypothesis that aberrant responses increase with the later location of items, two methods are used. First, the questionnaire was divided into two sections or three sections. Second, items were selected in order as moving windows. After each section was evaluated for the level of aberrant responses, the number of aberrant responses increased with the later location of items in the long questionnaire which it took about 50 minutes to be completed. But, respondents may not make as much an aberrant response about the long questionnaire compared to short questionnaire which took approximately 20 minutes to complete.

keywords : Markov Random Fields, Sequential Pattern Detection, Aberrant Response Detection

Student Number : 2012-22914